GRAIN SIZE MATTERS: L1 EFFECTS IN MORPHOLOGICAL, COMPLEXITY, ACCURACY, AND FLUENCY DEVELOPMENT

by

Hillary Schepps

B.A., Yale University, 2010

Submitted to the Graduate Faculty of

The Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Arts

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Hillary Schepps

It was defended on

March 5, 2014

and approved by

Dawn E. McCormick, Ph.D., Linguistics Department

Yasuhiro Shirai, Ph.D., Linguistics Department

Thesis Director: Alan Juffs, Ph.D., Linguistics Department

GRAIN SIZE MATTERS: L1 EFFECTS IN MORPHOLOGICAL, COMPLEXITY, ACCURACY, AND FLUENCY DEVELOPMENT

Hillary Schepps, M.A.

University of Pittsburgh, 2014

In second language acquisition (SLA), three constructs of complexity, accuracy, and fluency (CAF) have been isolated to evaluate learners' language performance and development (Brumfit, 1984; R. Ellis, 2008; Skehan, 1989, 1998). However, the emergence and interaction of these subsystems over time remain debated (Housen & Kuiken, 2009). This thesis examines whether learners follow a shared developmental path in SLA (Vercellotti, 2012), whether each learner follows her own unique, idiosyncratic path (Larsen-Freeman, 2006), as well as the role of a learner's first language (L1) in accounting for individual variation, especially in morphosyntactic accuracy (N. Ellis, 2006; Luk & Shirai, 2009). To explore these questions, this research analyzes the development of CAF in the semi-spontaneous spoken output of 30 learners of English (15 with L1 Chinese, 15 with L1 Arabic) over eight months as they progress from a low intermediate to a high intermediate level of proficiency while enrolled in an intensive English. I also consider their accuracy on six grammatical functors to examine L1 effects in morphological and syntactic development.

This research does not find a significant L1 effect in CAF development between-groups, but there is a reliable effect for the interaction between CAF and the L1, and overall, the Arabic learners have higher fluency and accuracy. Moreover, there was a significant interaction between time and gains in fluency, but only for the Chinese learners. In addition, there are clear L1 effects in grammatical functor accuracy scores, with Arabic speakers significantly more accurate than

the Chinese on plural *–s* at both levels and on third person singular present *–s* at the higher level. These results suggest that the grain size of measurement matters, because between-group L1 effects are only significant on the specific accuracy measures. It follows that learners' second language development is best operationalized by looking at global as well as specific measurements, as general measurements—too often employed in the literature—only tell a part of the story. Furthermore, this research confirms the observation that group averages tend to conceal significant individual variability (Skehan, 2009), and that L1 may not be the best grouping factor when employing global measurements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This research was supported by a research grant from the Pittsburgh Science of Learning Center to Dr. Alan Juffs, which is funded by National Science Foundation grant number SBE-0836012 (PSLC, http://www.learnlab.org). It was previously funded by NSF award number SBE-0354420.

I sincerely thank my thesis committee for their involvement and support. Dr. Dawn E. McCormick's pedagogical guidance was crucial to my implementing research and theory in my own classroom. Dr. Yasuhiro Shirai's knowledge of the literature was helpful in tracking down relevant research and publications. And finally, my thesis adviser Dr. Alan Juffs has earned my infinite gratitude for his patience, wisdom, and for showing me how to turn my research interests into potentially significant contributions to the field of Second Language Acquisition.

This thesis would not have been possible without the help of my classmates, ELI colleagues, professors, and family. In particular, I would like to thank my fellow graduate students Alexis Cherewka, Maritza Nemoga, and Zhaohong Wu for their patience and moral support. Ben Madore's help with the ELI database was critical, as was Christine O'Neill's guidance in tracking down ELI materials and RSA prompts from semesters past.

Finally, I would like to dedicate this thesis to my partner Roberto Roselli, whose encouragement and good humor made this work and my graduate career possible. As he often

reminds me regarding my graduate studies, “You wanted the bicycle; now pedal!” Let us consider this thesis the result of a journey on quite the bike.

Any errors in the thesis are my fault alone.

1.0 INTRODUCTION

The complexity, accuracy, and fluency (CAF) constructs comprise three perspectives from which to assess second language (L2) performance, both as indicators of learners' L2 proficiency underlying a given performance, and as a way to measure developmental progress over time (Housen & Kuiken, 2009). While Brumfit (1984) was among the first to distinguish explicitly between fluency- and accuracy-oriented activities in the classroom, Skehan (1989, 1998) was the first to define an L2 model using CAF as the three principal dimensions of proficiency.

In current research, complexity is seen as "[T]he extent to which the language produced in performing a task is elaborate and varied," (R. Ellis, 2003, p. 340); this research focuses on syntactic complexity by subordination (e.g., subordinate clauses such as verbal complements, relative clauses, etc.). Accuracy refers to the ability to produce error-free speech, with errors classified as syntactic, morphological or lexical deviations from target language (TL) norms. Yet the types of errors themselves are significant, with many attributable to the learner's first language (L1), and others having a universal basis (Corder, 1967). Finally, fluency refers to the ability to produce speech with minimal pausing, hesitation, reformulation, and self-correction (R. Ellis, 2003, p. 342).

1.1 BACKGROUND

The relationship between the CAF constructs is multifaceted and highly debated.[1] Some researchers argue that all three are interrelated and should increase hand-in-hand over time (Vercellotti, 2012). For example, Robinson's Cognition Hypothesis (2003, 2005) predicts that complexity and accuracy (in the form of grammaticization) will increase together during development as learners perform increasingly complex tasks (cf. Robinson, Cadierno & Shirai, 2009). In contrast, Skehan and Foster (1997, 1999) note learners' limited attentional resources and the tension between risk-taking and control in language production. Skehan and Foster (1999) hypothesize that "consistent prioritization of complexity might lead to a wide range of structures but a failure to move toward accuracy and control" (p. 97). Likewise, Larsen-Freeman (2006) argues that learner's focus on one measure (e.g., accuracy) will necessarily reduce his/her available attention to other measures. This notion based on limited attentional resources is better known as the Trade-off Hypothesis (Skehan, 1998), but it remains unclear how trade-offs manifest themselves over time.

1.2 STATEMENT OF PROBLEM

Past research has found conflicting results with respect to CAF development across learners over time. Larsen-Freeman (2006) analyzed five L1 Chinese learners of English of comparable proficiency performing the same oral and written tasks four times over a six-month period. She found a common tendency to improve over time, but variation between and within learners at

[1] Interested readers are urged to consult *Applied Linguistics*, Volume 30, Issue 4, which deals specifically with CAF.

both micro- and macro-levels of analysis despite their common L1. She argues that as each learner's unique path of development unfolds over time, it is constrained by the learner's allocation of limited attentional resources to each of the CAF subsystems. She situates her discussion in a dynamic systems theory framework, arguing for a position that views learners' performance from an emergentist standpoint and language as a complex adaptive system in which CAF constructs are certainly interconnected but do not necessarily correlate and certainly do not follow linear growth trajectories.

Using data similar to those used in this thesis, Vercellotti (2012) investigated 66 learners of L1 Arabic, Chinese, and Korean at varying levels of proficiency over the course of three to nine months. In contrast to Larsen-Freeman, Vercellotti found that learners followed the same growth trajectories in CAF development regardless of L1, and that nearly all CAF subcomponents exhibited linear growth. However, it is possible that these results are due to the combination of so many learners' data through the lens of a macro-analysis, which risks concealing significant individual variation both within and across performances. Furthermore, the lack of L1 effects may be due to Vercellotti combining all errors (lexical, syntactic, and morphological) and weighting them equally, without distinguishing between them and whether errors can be attributed to learners' L1.

1.3 RESEARCH QUESTIONS

This research attempts to reconcile Larsen-Freeman and Vercellotti's conflicting results about learner L2 progress over time (i.e., unique vs. shared developmental paths) by looking deeper at the role of L1 in the accuracy construct and performing an in-depth micro-analysis of accuracy

on six grammatical functors (plural *–s*; articles *a/an* and *the*; past regular *–ed*; irregular past; and third person singular present *–s*[2]) in addition to a macro-analysis of CAF performance over time. The research questions are:

(1) Do L1 and cultural background influence the development of CAF over time?

(2) Is learners' accuracy on six grammatical forms influenced by their L1?

Both questions will be investigated by looking at the semi-spontaneous oral production of 30 ESL students (15 with L1 Arabic, 15 with L1 Chinese) enrolled in an intensive English program as they progress from a low to high intermediate level of proficiency, with six data collection points over eight months.

Research question 1 is investigated by a mixed Repeated Measures Analysis of Variance (RM ANOVA) test with 2 (L1) x 3 (CAF) x 6 (time) that examines whether there are effects for these variables and their interaction. Research question 2 is explored via a 2 (L1) x 6 (grammatical functor) x 2 (time) RM ANOVA. In addition, two implicational scales are used to analyze the role of L1 in the emergence and accuracy of the six grammatical functors under investigation. These tables are also used to examine the degree to which there exists a "natural order" of emergence and accuracy of these functors, and how L1 influence can affect the order. Specific hypotheses for the research questions are presented in Section 4.5.1.

[2] Possessive *'s*, progressive *–ing*, copulas, and auxiliaries were also considered in this research, as they appear in other "morpheme order" studies, but they did not occur frequently enough in the corpus to allow a meaningful, systematic analysis.

1.4 SIGNIFICANCE

This study comes at a critical point in SLA research, as some applied linguists abandon notions of a one-size-fits-all developmental path characterized by fixed stages, such as those suggested by Processability Theory (Pienemann, 1998), and view learner development as the interaction between a number of dynamic systems that contribute to L2 as a complex adaptive system (Larsen-Freeman, 2006, 2009). The considerable variation observed between and among individual learners demands this kind of in-depth analysis of oral performance over time. In past research, individual differences in performance have often been obscured by macro-analyses that group learners together based on their L1 and find reliable differences in averaged performances over time. However, these macro-analyses may obscure individuals' performances and unique paths of development (Larsen-Freeman, 2006, p. 612), demanding the kind of microanalysis of individual developmental paths that this research employs in its accuracy measures.

The results of this study should also inform pedagogy. If it is found that learners favor one subsystem (e.g., accuracy) over others (e.g., fluency), ESL and EFL teaching techniques can be updated accordingly to favor a use of resources where no one aspect of the CAF constructs is emphasized at the cost of others, and all facets of communicative competence are addressed proportionally to the learner's objectives (Skehan, 1998).

1.5 THESIS OVERVIEW

The organization of the thesis is as follows. Chapter 2 reviews the Analysis of Speech unit, past research on CAF development, and how these constructs are operationalized in this research.

Chapter 3 discusses past research on a specific accuracy measure through the lens of the "morpheme studies," exploring both universal explanations and the role of the L1 in explaining the "natural order" of morpheme acquisition and accuracy. Chapter 4 presents the methodology of the current research, the participants, as well as research questions and hypotheses. Chapter 5 presents both descriptive and inferential statistics for the CAF and grammatical functor analyses, as well as the significance of these results. Potential implications are discussed in Chapter 6. Finally, Chapter 7 contains a conclusion, the limitations of this study, and directions for further research.

2.0 MEASURING THE COMPLEXITY, ACCURACY AND FLUENCY OF INDIVIDUAL PERFORMANCES

Complexity, accuracy and fluency comprise three perspectives from which to evaluate L2 performances. They are usually operationalized so that their measures take the form of ratios and frequencies. Norris and Ortega (2009) explain that the primary reason for measuring CAF in the area of instructed SLA research is to "account for how and why language competencies develop for specific learners and target languages, in response to particular tasks, teaching, and other stimuli, and mapped against the details of developmental rate, route, and ultimate outcomes" (p. 557). In other words, the results of such measures can inform not only the theories and mechanisms of SLA, but also the pedagogy that occurs in instructed SLA environments.

In the past, CAF measures have been applied to evaluate both oral and written performance. However, it has been noted that performance characteristics are often contingent on task type and demands, where the accuracy of certain structures and forms in written data may not reflect what the learner is capable of spontaneously producing because of Monitor effects (Krashen, 1985) and other factors related to style-shifting (Tarone, 1985). For example, Larsen-Freeman (1975, 1976) observed that learners' accuracy on grammatical functors varied with task type, with higher grammatical accuracy on certain forms in written than spoken tasks. In addition, Lardiere's (2007) case study of L1 Chinese learner Patty found that she used plural *–s* in about half of the obligatory spoken contexts, but 84% of the time in written contexts (p. 199).

Therefore, in order to attenuate Monitor effects, this thesis only considers semi-spontaneous oral output from a specific task called Recorded Speaking Activities (RSAs, see Section 4.1), following Spinner (2007, 2011) and Vercellotti (2012).

Skehan (1998) argues that the three CAF components draw on different types of linguistic knowledge. Rod Ellis (2008) elaborates on this idea, commenting that: "Fluency requires learners to draw on their memory-based system, accessing and deploying ready-made chunks of language and, when problems arise, using communication to get by" (p. 490). In other words, fluency relates to control over L2 knowledge, reflected in the speed and ease of accessing relevant L2 knowledge to communicate information in real time, "with control improving as the learner automatizes the process of gaining access" (Wolfe-Quintero, Inagaki, & Kim, 1998, p. 4). Meanwhile, complexity and accuracy reflect syntactic processing and thus draw on rule-based linguistic knowledge. Housen and Kuiken (2009) and Wolf-Quintero et al. (1998) maintain that a link exists between complexity and accuracy due to the state of the learner's interlanguage knowledge, which is partly declarative/explicit and partly procedural/implicit and may take the form of L2 rules or lexico-formulaic knowledge. Both complexity and accuracy relate to the representation of L2 knowledge and the level of analysis of internalized linguistic information (Housen & Kuiken, 2009, p. 462). Complexity and accuracy are often viewed as in competition with one another, because increasing complexity reflects risk-taking and restructuring of learner languages, while accuracy measures a learner's ability to control his/her existing resources and avoid errors. Skehan (1998) argues that the three CAF aspects of performance are at least somewhat independent from each other and are subject to different influences, whose exploration can have important pedagogical implications in a task-based learning framework (p. 5).

Since the 1990s, CAF measures "have appeared predominantly, and prominently, as *dependent variables* in SLA research" which emerge as "distinct components of L2 performance and L2 proficiency which can be separately measured and which may be variably manifested under varying conditions of L2 use, and which may be differently developed by different types of learners under different learning conditions" (Housen & Kuiken, 2009, p. 462, emphasis original). In this research, I consider CAF as dependent variables that may potentially be influenced by a number of independent variables including learner L1 and time (i.e., development). More recently, CAF have been the independent variables of SLA investigations, in which they are epiphenomena of the psycholinguistic mechanisms and processes that underlie L2 acquisition, representation, and processing (Lennon, 2000; O'Brien, Segalowitz, Freed, & Collentine, 2007; Segalowitz, 2007; Skehan, 1998; Tavakoli & Skehen, 2005, among others). However, such a perspective is inconsistent with my research goal of examining the role of L1 in CAF development; therefore, in this research, CAF are the dependent variables.

Each of the CAF constructs can be operationalized on global and specific levels to capture development in performances over time. However, before describing how CAF are operationalized in this thesis, I will first review the basic unit of measurement in spoken data, which is critical to the measures employed in this thesis: syntactic complexity by subordination, global clausal accuracy, and fluency speed.

2.1 BASIC UNIT OF MEASUREMENT: AS-UNIT

Before analyzing complex oral data on a micro-level, it is necessary to divide the data into well-defined, valid, and reliable units in order to measure aspects of CAF. Such measurements are

often expressed as the ratio of number of word or errors per clause, clauses per sentence, or are based on the internal complexity of a clause or utterance. Although written data is easily separated into sentences by punctuation, oral data presents complications. This is because native and non-native speakers do not speak in sentences, but in idea units (Luoma, 2004).

In the past, speech samples have been analyzed in terms of productivity, via mean length of utterance (MLU, cf. Brown, 1973; Hakuta, 1974), and complexity. However, MLU and other productivity measures alone may be insufficient to reflect cognitive processes like syntactic processing, as a highly productive L2 learner's output may exhibit low syntactic complexity because of a reliance on "chunks" or memorized phrases (Foster, Tonkyn, & Wigglesworth, 2000). Therefore, it is necessary to measure not only productivity but also complexity to get a fuller picture of L2 development.

Critically, the unit used to operationalize complexity must satisfy the well-established methodological criteria of validity and reliability. As Crookes (1990) notes, in order for the unit to be psycholinguistically valid, it must reflect a psycholinguistic planning process. Foster, Tonkyn, and Wigglesworth (2000) distinguish between macro-planning, i.e., multi-"sentence" stretches of speech, and micro-planning, which is associated with shorter units at the clause or sentence level (p. 355). The authors predict that from a planning perspective, increases in proficiency are indicated by the ability to keep track of more micro-units, thus allowing the speaker to communicate a more complex message in a shorter time span. Foster et al. summarize how the unit should reflect "what the performer can achieve in a single chunk of micro-planning activity, and how particular types of plan may affect the complexity, accuracy, and fluency of the language that is produced" (p. 356). In addition, this unit of measurement should be applicable crosslinguistically, across data sets from both native and non-native speakers. It therefore

requires a clear definition and examples of application in order to be used reliably in speech analysis.

Prior definitions of units of measurement have been vague, with few published examples containing real data and consequently, limited applicability of such ill-defined units. Prior semantic measures have included the proposition, C-unit, and idea unit, but are difficult to apply, as ideas and arguments are rarely clear-cut, and it is impossible to establish boundaries of ideas with any certainty. Intonational measures include the tone unit/phonemic clause and the utterance, which occurs under a contour, is bounded by pauses, and contains one semantic unit (Crookes, 1990, p. 187). However, intonational units are only reliable for native speakers, as L2 data is often marked by unnatural, mid-clause pauses (i.e., not at unit boundaries) often resulting from message formulation or lexical search (Skehan & Foster, 2008). L2 data would thus contain more tone nuclei than a native speaker producing the same message. Furthermore, low proficiency learners' L2 data may also exhibit such high levels of dysfluency that intonational criteria are not applicable whatsoever.

Syntactic units of measurement have fared best because they are easier to identify objectively than semantic or intonational units. One example is the clause, or S-node, identified as "either a simple independent finite clause, or a dependent or non-finite clause" (Foster & Skehan, 1996, p. 310). Other units at the supra-clausal level include the sentence—problematic for spoken data for obvious reasons—and the t-unit, defined by Hunt (1970) as "a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it" (p. 4). Foster et al. (2000) argue that supra-clausal units "allow the analyst to give credit to performers who can embed clauses and hence construct chunks of speech which reflect more sophisticated planning processes" (p. 362). However, definitions like the t-unit are insufficient given the

elliptical nature of speech—a one-word response to a question does not constitute an independent clause. Therefore, it is necessary to define not only a supra-clausal level unit, but also the clausal and subclausal constituents that comprise it. Despite clear ideas about what the unit should measure, structures including *because* adverbial clauses, coordinated phrases, topical noun phrases, as well as features of scaffolding and interruption, have been notoriously difficult for analysts to parse given the ambiguity of whether such structures comprise their own unit. Based on the lack of a well defined and reliable unit of measurement for spoken data, Foster et al. (2000) propose a mainly syntactic Analysis of Speech unit (AS-unit) in their article "Measuring Spoken Language: A Unit for All Reasons." They define the AS-unit as "a single speaker's utterance consisting of *an independent clause, or sub-clausal unit*, together with any *subordinate clause*(s) associated with either" (p. 365, emphasis original). An independent clause will minimally consist of a clause including a finite verb, while a subordinate clause consists minimally of a finite or non-finite verbal element, plus at least one other clause element such as a subject, object, complement, or adverbial. Thus, "I have no chance to visit" consists of one clause, while "I have no chance :: to visit you" consists of 2 clauses, where *you* is the other clausal element in the verbal complement. Consider the following examples of the various functions of subordinate clauses, where :: marks clause boundaries, | marks AS-unit boundaries, and parentheses surround pauses in seconds.

Table 1. Subordinate clause functions in AS-units

	Function of the subordinate clause	Example (from Foster et al., 2000, p. 367-8; pauses are mine)	Clauses	AS-units
a	subject (initial or postponed)	| sometimes it creates problems :: that he knows nothing |	2	1
b	verb complementation (object)	| and er they told :: that there was no food crisis |	2	1
c	complement	| I wish :: to er visited other areas of English |	2	1
d	concatenative verb complementation	| I would like :: to ask you :: if you can give me three weeks leave now |	3	1

e	phrasal post-modifier	| still in our country the school and er college students learner the English :: which were er taught to the students before thirty years |	2	1
f	adverbial	| when I was in the university :: er I have specialized in this er subject |	2	1
g	adverbial, 1 tone unit	| and I can bring him tomorrow together :: where you can talk with him |	2	1
h	adverbial, 1 tone unit	| I can understand (0.4) :: when I read scientific English |	2	1
i	adverbial, 2 tone units	| specifically for reading scientific papers | because er all the papers that er arrived to the library in Chile (0.5) :: are English paper |	3	2

Importantly, if the adverbial follows the main clause as in examples (g), (h), and (i), it must be in the same tone unit as the main clause in order to be included in the preceding AS-unit. In the case of coordinated phrases, the VPs will generally belong to the same AS-unit unless the first phrase is marking by falling or rising intonation, and is followed by a pause of at least 0.5 seconds (p. 367). Finally, topicalized noun phrases belong to the unit of which they are the topic; however, if the NP is marked by falling intonation and a pause of at least 0.5 seconds, then the NP comprises a separate AS-unit (p. 369).

Foster et al. also address features of dysfluency including false starts, repetitions, and self-corrections. They define a false start as "an utterance which is begun but then either abandoned altogether or reformulated together in some way" (p. 368) and only count it as an AS-unit if the utterance produced before the abandonment meets AS-unit criteria, with the remainder recorded as a false start, here indicated by { }. Consider the following example of a false start in Arabic learner 241's data:

(3) A 241:[3] |{we have to} first of all we have :: to eat together (0.9) | and sit in the floor : :: to make a big circle |

[3] The notation adopted for learner ID is a letter followed by a number. The letter, A or C, denotes L1 Arabic or Chinese, while each number was independently assigned by the ELI database.

Note that the second AS-unit is a coordinated phrase, but because it is offset but a pause of at least 0.5 seconds and comprises its own tone unit, it is counted as a unique, non-subordinate AS-unit.

Repetitions occur when the speaker repeats previously produced speech, yet it is necessary to indicate those which occur to allow time for online planning, as in example (4), and those which are rhetorical in nature, as in (5):

(4) C 126: | {in (0.6) in my} in my country now many people have pets |

(5) A 12: | this job was very very good salary job |

Only rhetorical repetitions are considered part of the AS-unit; the others are not counted in a measure of length (e.g., syllable or word count) of the unit.

Next, Foster et al. define a self-correction as "when the speaker identifies an error either during or immediately following production and stops and reformulates the speech; self-corrections will therefore include an element of structural change" (p. 368). They only count the final version/formulation as an AS-unit, with previous ones excluded, as in the following examples (6) and (7):

(6) A 29: | and there is {more two bigger} two big city :: {which is} which are Jeddah and Ademann |

(7) C 282: | {when I work} when I worked in Taiwan :: the important person is the one :: {who is my} who was my boss |

Foster et al. advise not including false starts, functionless repetitions, and items that are replaced for grammatical or lexical reasons in a word or syllable count when measuring the length of AS-units (fn. 10, p. 374). This is especially relevant for my fluency rate measure of words per minute, because characteristics of dysfluency such as false starts and functionless repetitions should numerically reflect a drop in fluency and do exactly that via fewer words per

minute. Overall, because Foster et al.'s (2000) comprehensive definition of their AS-is so "accessible, clearly defined, and easily applied," (p. 371), I employ it in this research.

2.2 COMPLEXITY

Among the three CAF constructs, complexity is likely the hardest to define and operationalize since it is "the most complex, ambiguous, and least understood dimension of the CAF triad" (Housen & Kuiken, 2009, p. 463). For one, it tends to be conflated with development or growth in SLA. Consider Skehan and Foster's (1999) definition of complexity as

> [T]he capacity to use more advanced language, with the possibility that such language may not be controlled so effectively. This may also involve a greater willingness to take risks, and use fewer controlled language subsystems. This area is also taken to correlate with a greater likelihood of restructuring, that is, change and development in the interlanguage subsystem. (p. 96-97)

Pallotti (2009) notes that if complexity is considered only as "more advanced" or "challenging language," then it is not a descriptor of a given performance but necessarily compares it to previous performances, indicating progress. Rod Ellis (2003) considers complexity as "elaborate and varied" language (p. 340), but this notion can refer to both lexical and syntactic aspects of linguistic complexity. In the current research, I limit my analysis to syntactic complexity by subordination, as lexical variety and the variety of syntactic structures employed will likely be too contingent on the speech topic, which varied across learners.

2.2.1 Operationalizing syntactic complexity

Norris and Ortega (2009) reviewed the main measures of syntactic complexity used in 16 recent task-based SLA studies and call for a more organic approach to measuring CAF. Some measures of overall or general complexity are based on length and divide the number of words, morphemes or characters by a specific production unit (e.g., MLU). The unit in the denominator may be single- or multi-clausal. If it is a multi-clausal unit, then an increase in the length measurement may be due to subordinate clauses, adjectives, prepositional phrases, or nonfinite VPs that modify other elements through non-subordinating clauses, etc. (Norris & Ortega, 2009, p. 561). For this reason, it is a global measurement, as it does not specify the source of the complexity. In contrast, if the length measurement contains a clausal unit in the denominator, it can indicate a different type of complexity, unaffected by subordination. Norris and Ortega explain that when this measurement increases, it can only be due to pre- or post-modification within a clause, the use of nominalizations, or the reduction of clauses into phrases (p. 561). Thus, this measure of subclausal complexity only measures phrasal elaboration. But as stated earlier, length measurements alone may be misleading, as some learners tend to rely on memorized phrases or chunks that do not to require syntactic processing; therefore, high overall length scores may not genuinely reflect the level of syntactic complexity or analysis underlying a given utterance.

Although global measurements of average length do not give insight as to how the lengthening was achieved, length is considered to be the best measurement of proficiency in SLA writing (Larsen-Freeman & Long, 1991). However, in oral output, some length-based complexity measures load highly with fluency measures (Norris & Ortega, 2009), so much so that some researchers have used length-based measurements as a measure of fluency (e.g.,

Larsen-Freeman, 2006). Therefore, lest the syntactic complexity and fluency constructs be conflated, it is necessary to use not the utterance in the denominator but a unit determined by syntactic criteria such as the AS-unit. Norris and Ortega also advise measuring complexity in different ways.

Other measures of syntactic complexity consider the degree of subordination, calculated by adding up the number of clauses in the numerator and dividing by the production unit, as suggested by Foster et al. (2000). This has been used frequently in SLA research for segmented oral data. The value increases when more subordinate or dependent clauses are used in an AS-unit, which captures a very specific type of complexity common in intermediate level L2 learners' output. Finally, Norris and Ortega referred to measures of variety, sophistication, and acquisitional timing of grammatical forms such as Pienemann's Rapid Profile (1998) assessment system. However, the Profile is not suited for this thesis, given the variety of RSA topics and the diverse tenses and structures that they elicit. For example, when Spinner (2011) performed an analysis on RSAs from the same corpus as this research, she had to exclude certain elements from the Profile analysis because not enough tokens were produced, even when data from multiple observations were collapsed together.

Norris and Ortega stressed the importance of combining different sub-constructs to capture distinct dimensions of syntactic complexity. This is exactly what Vercellotti (2012) did, by measuring global complexity in mean number of words per AS-unit; subclausal complexity in mean number of words per clause; and complexity by subordination in a ratio of clauses per AS-unit (p. 73). She found that the first two measures (AS-unit length and clause length) were highly correlated, suggesting that they are tapping into the same construct.

2.2.2 Measuring syntactic complexity by subordination

Because an investigation of each of the dimensions of complexity is beyond the scope of the statistics available for this thesis, I limit my measurement to syntactic complexity by subordination in terms of clauses per AS-unit. This measure was chosen based on the way learners' style of expressing complex ideas changes over L2 development. Norris and Ortega (2009) note that at beginning levels, coordination is predicted to be the most likely source of syntactic complexity (as argued by Bardovi-Harlig, 1992); while at intermediate levels, it will be accomplished by subordination. At advanced levels, learners should rely more on phrase-level complexification to produce denser sentences and achieve syntactic complexity.[4] For this reason, Norris and Ortega suggest that "subordination should be a useful and powerful index of complexification at intermediate levels" (p. 563) – exactly the level of the learners under investigation in this thesis. Thus, I operationalize syntactic complexity by subordination as the mean number of clauses per AS-unit. Although Vercellotti (2012) recommends that future studies measure not the number of clauses per AS-unit but the ratio of *finite* clauses to AS units (p. 153), I do not adopt this measure as it would limit the comparability of the current thesis to other research using the AS-unit as defined when published.

[4] Norris and Ortega's claims are based on Halliday and Martin's (1993, p. 31-41) proposal that over development learners shift from a dynamic style of expression, characterized by increasing subordinate clauses, to a synoptic variety that relies on phrasal elaboration.

2.3 GLOBAL ACCURACY

Accuracy is measured in terms of whether learner interlanguage corresponds to target language (TL) norms. Deviations from TL norms may take the form of errors or mistakes. Corder (1967) defines errors as deviations that arise because of a lack of knowledge or a breakdown of competence, while mistakes are performance phenomena that reflect "processing failures that arise as a result of competing plans, memory limitations, and lack of automaticity" and are also present in native speakers' speech (R. Ellis, 2008, p. 48). Corder argued that only errors, not mistakes, should be analyzed, but it is difficult to make this distinction in the data. Furthermore, this distinction ignores variability, as competence is not homogenous but heterogeneous. Because some learners might sometimes use a target form correctly and other times not, one cannot necessarily conclude that the learner "knows" the target form, nor that deviations from the target form represent errors. Instead, the degree of accuracy often depends on context, as is the nature of emergence of forms (R. Ellis, 2008).

When identifying errors, it is necessary to distinguish which TL norms are relevant, as forms that might be acceptable in one variety of English (e.g., "I like sport" in British English) are considered errors in other varieties. For this reason, I chose American English as the TL since students are enrolled in an instructed SLA environment at an intensive English program in Pittsburgh, PA. Similarly, it is necessarily to distinguish between well-formed utterances that adhere to TL norms, and those that are superficially "accurate" but do not mean what the learner wants them to mean. For example, Rod Ellis (2008) mentions the example of a learner producing the well-formed sentence "The wind was stopped" when she meant "the wind stopped." In order to deal with such utterances, I rely on Lennon's (1991) definition of an error: "A linguistic form

or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speaker's native speaker counterparts" (p. 182).

2.3.1 Types of errors

Errors in this thesis are categorized according to Dulay, Burt, and Krashen's (1982) typology and may be lexical, syntactic, or morphological. Pronunciation problems are not considered errors unless the output comprises non-existent words where the intended word is unintelligible. Consider some examples of morphological vs. syntactic errors categorized according to Dulay et al.'s schema (1982, p. 148-150):

Table 2. A sample linguistic category error taxonomy

Linguistic Category	Error Type	Example
Morphological	wrong indefinite article	he is **a*** important person
	omitted plural *–s*	he got some **leaf***
	omitted possessive *'s*	in my **country*** history
	omitted third singular *–s*	he **live*** in Taiwan
	omitted past reg. *–ed*	I **assist*** him for many years
	oversupplied past reg. *–ed*	he **choosed*** his brother
	simple present form for irregular past	she **take*** me everywhere last year
	comparative construction *more + er*	the spa is **more*** cheaper
Syntactic	NP: omitted article	he studies at * University of Pittsburgh
	NP: omitted subject pronoun	I think * is a good idea
	VP: omitted copular	he * in the water
	VP: omitted auxiliary	I * studying at the ELI
	VP: oversupplied auxiliary	it **was*** happened last year
	VP: oversupplied prog. *-ing*	I **seeing*** him every weekend
	VP: subject-verb agreement	we **was*** friends
	word order	he drinks **often*** the beer

Lexical errors, which are not illustrated in Table 2, may include the wrong content word, the wrong function word (preposition, etc.), the wrong part of speech, etc. See Section 4.4.2.1 for examples of errors in this corpus and how they were coded.

Accuracy can be influenced by a number of factors. Tarone (1985) distinguished between attention to speech and redundancy of elements, that is, whether elements have discourse significance and encode important aspects of meaning. She found that discourse salience for forms such as articles and object pronouns causes them to be used more accurately in narratives and interviews than other contexts where meaning is less critical. According to Skehan (1998), such a result suggests that attention paid to speech (usually ensuring higher accuracy) must be interpreted "by saying that discourse processing requires particular attention to be paid to forms which have unusual salience in communication" (p. 67). See Chapter 3 for an exploration of the accuracy of such forms.

The accuracy of a given performance may also be affected by the complexity. Skehan and Foster (1999) note that higher accuracy may reflect both "higher levels of control in the language, as well as conservative orientation, that is, avoidance of challenging structures that might provoke error" (p. 96), implying a tension between control/accuracy, on the one hand, and risk-taking/restructuring on the other; see Section 2.5 for an exploration of tensions and trade-offs.

2.3.2 Measuring accuracy in error-free clauses

In this research, global accuracy is operationalized as the number of error-free clauses divided by the total number of clauses in a given RSA performance. This measure was chosen because a clause is a psychologically valid micro-planning unit, and is reliable across performances and learners (Foster et al., 2000). While Vercellotti (2012) measured not only clausal accuracy but also AS-unit accuracy (error-free AS-units divided by total AS-units), she found that the two measures were strongly correlated and therefore suggests, "for many studies, clause level

accuracy might suffice" (p. 153). Furthermore, learners at a low intermediate level of proficiency struggle to produce completely error-free AS units, especially as the length of the unit increases. Therefore, AS-unit based accuracy measures may be "too demanding" at this level (Vercellotti, 2012, p. 169) and risk floor effects and a consequent type II error. The smaller grain size of clause instead of AS-unit will allow learners to earn some "partial credit" if a clause but not the entire AS-unit in which it is occurs is error-free.

Nevertheless, this global accuracy measure conceals the source of errors, as each one is equally coded whether it is due to L1 effects, universal processing pressures, or individual idiosyncrasies—a common problem in CAF research that employs excessively general measures of constructs. For this reason, in addition to global accuracy, I also look at specific accuracy on six grammatical functors. See Chapter 3 for an in-depth justification of this accuracy measure, inspired by the morpheme studies (Bailey, Madden, & Krashen, 1974; Dulay & Burt, 1973, 1974; Krashen, 1977; among others).

2.4 FLUENCY

Among all three constructs, fluency is the easiest to conceptualize but the hardest to operationalize because it is so multi-faceted. Fluency can be considered the ability to speak quickly, eloquently, or smoothly (see Housen & Kuiken, 2009) but on closer inspection, this notion is insufficient given the multi-dimensional nature of fluency.

2.4.1 Types of fluency: Breakdown, speed, and repair

Skehan (2003) and Tavakoli and Skehan (2005) differentiate between three aspects of fluency: breakdown fluency, speed, and repair fluency. While breakdown fluency is a silence-related metric measured by the length and number of unfilled pauses and silence, speed is a time-related metric, measuring the rate of speech by the number of syllables produced in a given time. Repair fluency is instead gauged by self-correction and can be evidenced by the frequency of reformulations, replacements, hesitations, and false starts. Importantly, these three subcomponents may differ by learner. For example, Tavakoli and Skehan's (2005) factor analysis revealed a difference between repair fluency and breakdown fluency among learners. Similarly, someone with a high rate of speech (i.e., high fluency speed) evidenced by a large number of words per minute may have high breakdown fluency because he/she is constantly performing reformulations and self-corrections. In this sense, the high frequency of "time-creating devices" such as fillers, hesitations, and re-phrasings may actually reduce the density of the information transmitted in speech in a given time frame and thus have the linguistic effect of slowing down time (Skehan, 1998, p. 31) and reducing overall fluency.

2.4.2 Measuring fluency in WPM

Generally, the ability to speak fluently is distinct from the ability to accurately use complex forms and structures (R. Ellis, 2008, p. 492), so it is critical to avoid conflating complexity and fluency. In order to do this, I consider fluency as a measure of speed: number of words produced per minute (WPM). Although this is less precise than syllables per minute given that a single word may have one or six syllables, for example, this decision leads to easier coding and it is

assumed that the differences between word length by syllables will "even out" in the tallies. Furthermore, my WPM measure combines the three subcomponents of fluency individuated by Skehan (2003). For example, if a learner has high breakdown fluency and pauses often, inclusion of the pauses in the WPM denominator will reflect the drop in fluency. Likewise, if a learner has high repair fluency and exhibits frequent reformulations, self-corrections, hesitations, repetitions, etc., then only the final formulations of these elements of the utterance are included in the final word count (following Foster et al., 2000, p. 374), and the exclusion of prior formulations will thus also reflect repair fluency. Although my measure of fluency in WPM does not distinguish between these three subcomponents of fluency in the final ratio, all three are still included in the ratio. In sum, this global measure of speech rate also encompasses pauses and repair because the more of these there are, the lower the total fluency WPM measurement will be.

2.5 RELATIONSHIP BETWEEN COMPLEXITY, ACCURACY AND FLUENCY

More interesting than the CAF measures alone is the extent to which these three dimensions interact in L2 performance and development (R. Ellis, 2008; Skehan, 2009).

Researchers who believe that learner language performance is necessarily constrained by limited attentional resources and processing capacity predict that a higher performance in one of the CAF components is associated with a lower performance in another (i.e., trade-off effects). Skehan (1998, 2009) and Skehan and Foster (1997, 1999) note the competition between meaning and form and predict that if learners emphasize meaning (via fluency) over form, then accuracy might suffer, while accuracy in turn competes with complexity via the tension between control and risk-taking/restructuring of interlanguage. Learners may consciously or unconsciously

emphasize one of the dimensions at the cost of others, be it due to task demands, individual preferences, or more likely, both factors.

While many researchers believe that CAF are in competition with each other, others maintain that they are "connected growers" and should increase together. For example, Robinson (2003) argues that learners can simultaneously access multiple attention pools that are not in competition with each other. According to his Cognition Hypothesis, thanks to attention control and interference, the complexity and accuracy dimensions of the CAF triad can simultaneously improve as task complexity is manipulated to increase the cognitive demands of the task. In other words, as learners attempt to produce the language that is required by the greater conceptual demands in the relatively increased complexity of the task, their language performance should improve developmentally through increasingly accurate grammaticization over time.

In contrast to the trade-off hypothesis and connected growers views of CAF inter-relationships, other applied linguists have adopted a dynamic systems theory approach to language performance, characterized by cognitive resources that are limited but connected to each other, and thus potentially compensatory. Every element in the system is interconnected, so changing any feature or variable will necessarily affect all others (de Bot, 2008). A dynamic systems approach obviates cause-and-effect models of language learning and rejects the notion that language growth is linear.

Let us now consider the potential interdependency of the CAF measures by looking at the various types of inter-relationships in greater detail.

2.5.1 Cross-sectional vs. longitudinal CAF studies

Before looking specifically at the results of past CAF studies, it is worth noting the difference between those that are cross-sectional and longitudinal. Cross-sectional studies look at learners' performances at one point in time. These studies, usually containing aggregated data across proficiency levels, often find that learner language performance might exhibit to trade-off effects. However, Larsen-Freeman and Long (1991) question the assumption that the aggregated data of learners of different proficiencies is comparable to longitudinal data. Furthermore, such cross-sectional studies say nothing about development. Norris and Ortega (2009) therefore suggest investigating how the CAF constructs interact in longitudinal studies to examine the process of language learning over time. Pallotti (2009) urges researchers to make a distinction between CAF measures that refer to properties of language performance as a product at one point in time, and linguistic development, which refers to a process characterized by the sub-dimensions of route and rate of acquisition. In Pallotti's words, "CAF measures can empirically be related to developmental dimensions, but the latter should not be analytically considered part of the former" (p. 594).

2.5.2 Trade-off hypothesis

The trade-off hypothesis (cf. Skehan, 1998) is based on the notion that learners have limited attentional resources (i.e., limited attentional capacity and working memory (Skehan, 2009)) and consequently will have a difficult time focusing simultaneously on all three aspects of production, prioritizing one with trade-offs in the others. Trade-offs are most likely to be evident, for example, if learners focus on complexity by taking risks with new complex structures, and

their accuracy and/or fluency suffers as they have less control over the language they are producing. Similarly, R. Ellis (2008) notes that an increase in fluency in SLA may occur at the expense of development of accuracy and complexity because of the differential development of knowledge analysis and automatization in SLA. Effects of the competitive relationship between CAF may be evident even in cross-sectional studies.

Research on CAF trade-offs has been popular since the 1990s, as applied linguists tried to attribute such trade-offs to factors such as task type, task demands, planning time, processing load, etc. For example, Skehan and Foster (1997) looked at CAF measures on three different tasks (personal, narrative, decision-making) and the effect of planning (unplanned, one minute planning, ten minutes planning) via a factor analysis. If learners' L2 proficiency were the key factor, then all CAF measures should load on this, but such was not the case. If task were the key factor, then a three-factor solution could be expected, with the three CAF measures each showing groups of loadings on the different tasks. Although there were some effects in that the narrative task was least accurate, complexity and fluency scores diverged across the two studies. In terms of planning time, Foster and Skehan found a monotonic relationship where greater planning resulted in higher complexity and fluency. However, accuracy was highest on the one-minute planning condition, showing trade-off effects in the allocation of attention. Overall, results of the factor analysis suggest there is a CAF trade-off, with a three-factor solution corresponding to these constructs occurring (i.e., the measures of each construct loading on the same factor). Thus, the three measures of performance are distinct and even enter in competition with one another. This led Skehan (1998) to conclude, "trade-off effects are operating very strongly. To improve in one area seems to be at the expense of improvement elsewhere. Selective rather than across-the-board improvement seems to be more realistic" (p. 112). However, such a conclusion must be

taken with a grain of salt, as the 1996 and 1997 studies were cross-sectional and any notion of "improvement" is highly speculative when applied to development.

Skehan and Foster (1999) later looked at the influence of task structure and processing load on narrative retelling performances in another cross-sectional study. They found that fluency was strongly influenced by the degree of inherent task structure, with more structured tasks resulting in higher fluency. In contrast, complexity of the language was affected by processing load, where higher processing requirements resulted in less complex language. Finally, performance accuracy was dependent on the interaction of task structure and processing load. However, the lack of correlations between fluency and complexity as well as between fluency and accuracy again suggests the independence of these domains, with a tension between meaning (fluency) and form (accuracy and complexity) in learner language. Moreover, the results of this study suggest that ESL teachers can manipulate tasks in order to encourage emphasis on specific CAF dimensions. For example, if fluency is the goal, then teachers should design tasks that are clearly structured. If accuracy is to be emphasized, then the authors suggest combining structured tasks with delayed processing conditions (p. 117).

Beyond the macro-level tension between meaning and form individuated by Skehan, Ahmadian (2011) also found a lower-level tension within form between complexity and accuracy. Based on the notion of limited attentional resources and Levelt's (1989) model of speech production, Ahmadian performed a longitudinal study to investigate the effects of massed-task repetitions on CAF, and whether improvement gained from task repetition transfers to a new task. He found that the experimental group of EFL learners who repeated a task 11 times every two weeks outperformed a control group on a new task in complexity and fluency but not accuracy. His research is situated within Levelt's (1989) model of speech production,

which differentiates between three overlapping stages: *conceptualization*, during which intentions and relevant information to be conveyed are selected and prepared in the form of a pre-verbal message; *formulation*, when conceptual representations are translated into linguistic structures; and *articulation*, during which the linguistic structures are transformed into actual speech. According to Ahmadian, when a learner repeats a task multiple times, attentional resources are freed up from conceptualization and may be allocated to different dimensions of oral performance including the formulation and articulation stages. The fact that the learners performing the repeated task improved in complexity and fluency but not accuracy on a new task again suggests a degree of competition between the three CAF dimensions and an inability to attend to all aspects of performance simultaneously. Furthermore, following Rod Ellis (2009), Ahmadian's research suggests that Levelt's speech production model, although initially designed for L1 speech, is also applicable to L2 learners and data.

To summarize, research on CAF trade-offs has found not only competition between meaning and form but also between the two sub-dimensions of form, complexity and accuracy, especially as related to task demands. Relying on Levelt's model, Skehan (2009) predicts that if the task demands lead learners to pay a higher amount of attention to content (i.e., conceptualization), then such a task will deplete the attention available for formulation, which is responsible for constructing structurally complex speech, resulting in a lower complexity score. However, Skehan notes that if higher conceptualization efforts do *not* interfere with the formulator, they will allow for more complex and more accurate language on difficult tasks, which brings us to the notion of connected growers.

2.5.3 Connected growers

Robinson's (2003, 2005) research is based on a multiple resources view of processing in which structural accuracy and functional complexity are not in competition but are instead connected growers. He distinguishes between the cognitive/conceptual difficulty of task designs that can either direct resources, aiding the performance, or disperse them, hindering it. Robinson's Cognition Hypothesis (2003, 2005) claimed that pedagogic tasks should be sequenced for learners in an order of increasing cognitive complexity, and that along resource-directing dimensions of task demands, increasing effort at conceptualization promotes more complex and more grammaticized L2 speech production. In other words, increasing the cognitive complexity of tasks will lead to greater linguistic complexity and accuracy over time.

In order to investigate this hypothesis, Robinson, Cadierno and Shirai (2009) focus on the impact of task properties on learners' L2 performance relying on two studies related to tense-aspect morphology when referring to time (Shirai, 2002) and lexicalization patterns when referring to motion (Cadierno, 2008). The authors increased the complexity of task demands in these two conceptual domains (time and motion) using specific, not general, measures of linguistic complexity and accuracy. They found that in more conceptually demanding tasks, there is more complex, developmentally advanced use of tense-aspect morphology than in less demanding tasks. Furthermore, there was also a slight trend to produce more accurate, target-like use of lexicalization patterns for referring to motion on increasingly complex tasks, but only for the L1 Danish and not the L1 Japanese learners studied by Cadierno. These results provide some (but not unequivocal) support for the Cognition Hypothesis. The pedagogical implication is that instructors should sequence tasks in increasing cognitive complexity if learners are to improve

simultaneously in grammatical complexity and specific accuracy. However, other support for the Cognition Hypothesis is less prevalent in the literature.

Vercellotti (2012) examined CAF development in the oral production data of 66 learners of English with L1s Arabic, Chinese, and Korean in an instructed ESL environment. She measured syntactic complexity in terms of mean length of AS-unit, mean length of clause, and subordination via clauses per AS-unit, while lexical variety was measured by a D-score (Malvern & Richards, 1997). Accuracy was measured by the proportion of error-free clauses and error-free AS-units. Finally, fluency was measured by a phonation time ratio (speaking time, excluding filled pauses, divided by total time), mean length of pause (average length of filled and unfilled pauses >200 ms), and mean length of fluent run (average number of syllables in an utterance bounded by pauses >200 ms).

Relying on hierarchical linear modeling, Vercellotti (2012) found that growth trajectories were the same for all measures, and that only lexical variety and mean length of fluent run measures exhibited non-linear growth. Vercellotti did not find trade-off effects in this longitudinal data, even though within-individual and between-individual correlations were also calculated. Vercellotti's findings lead her to conclude that instructed L2 performance growth is uniform, rather than along individual paths (in contrast with Larsen-Freeman (2006); see Section 2.5.4), and that L1 does not play a significant role in determining paths of development. Moreover, she argues that all subcomponents of CAF are connected growers and improvement in any one area does not imply a trade-off in another measure. However, as mentioned in my introduction, it is possible that such results are due to the combination of a large number of learners' data through the lens of a macro-analysis, which risks concealing significant individual variation both within and across performances. Furthermore, Vercellotti's lack of significant L1

effects may be due to her combination and equal weighting of all errors (lexical, syntactic, and morphological).

Besides a handful of studies, there is limited support for the Cognition Hypothesis and connected growers notion of CAF measures' development over time. Because positive correlations between complexity and accuracy are less frequent in the data than trade-off effects between these two facets of form, Skehan (2009) proposes interdependency between complexity and accuracy. If complexity and accuracy increase together, Skehan argues that it is not due to task difficulty, as Robinson's Cognition Hypothesis would predict. Instead, simultaneous growth reflects the joint operation of separate task and task condition factors. For example, the task structure may aid accuracy while the information manipulation during the task requires learners to use subordinate structures, leading to an increase in grammatical complexity. Finally, Skehan also argues that group data analyses may hide the fact that some individuals attend to one aspect of CAF while others attend to a different area. In other words, aggregated group data seem suggestive that two areas in a competitive relationship are improving when really, different learners are doing different things. For this reason, Skehan therefore suggests running correlations not only at the group level but also on individual performances. However, even when Vercellotti (2012) did exactly this, she did not find significant differences between individuals.

2.5.4 Dynamic systems theory

Some researchers reject both the trade-off hypothesis and notions of connected growers and instead adopt a different theoretical framework to account for CAF development over time. One such framework is Dynamic Systems Theory (DST), which de Bot (2008) employs to explain L2

development as a dynamic process of systems changing over time. Also known as chaos/complexity theory, DST is characterized by Complex Adaptive Systems (CAS).

Larsen-Freeman (1997) was among the first applied linguists to apply DST to SLA by arguing that language has all the characteristics of a dynamic system. She argues that like other CASs, language is dynamic because it changes over time both synchronically and diachronically. Language is also complex because it consists of different subsystems (syntactical, phonological, lexical, etc.) that interact and may amplify or compete with each other's effects (N. Ellis & Larsen-Freeman, 2009, p. 16). A change to any one variable in the system will necessarily result in effects on the other variables due to the inter-connected nature of complex systems. Next, language develops non-linearly and is often unpredictable and chaotic, due to a discrepancy between input and effects (de Bot, 2008, p. 167). More specifically, learners may exhibit backslides, stagnations, and jumps, and it is not clear which instances of instruction or input lead to which instances of learning (see Verspoor, Lowie, & de Bot, 2007 for further discussion). In addition, like other dynamic systems, Larsen-Freeman argues that an L2 is sensitive to initial conditions, as well as being open, self-organizing, feedback-sensitive, adaptive, and sensitive to attractors in development—all traits that characterize CASs.

Larsen-Freeman (2006) adopted a dynamic systems framework as she explored the development of five high intermediate L1 Chinese learners of English performing the same written and spoken tasks four times over a six-month period. She performed a macro-level analysis of the written data relying on the following quantitative measures: fluency, measured in the average number of words per t-unit;[5] grammatical complexity, via clauses per t-unit;

[5] As discussed earlier, Larsen-Freeman's length-based operationalization of fluency must be "handled with care" as it risks conflating fluency with complexity.

accuracy, through the proportion of error free t-units to total t-units; and vocabulary complexity, measured by a sophisticated type-token ratio that controls for length. Larsen-Freeman also employed qualitative measures by comparing the oral data organized into idea units to see how the language changed with each telling of the story due to potentially different distributions of cognitive resources.

In terms of the quantitative aspects of her study, Larsen-Freeman found that, overall, each construct showed growth in the group averages. However, the group averages conceal a great degree of individual variability, indicated by high standard deviations. Moreover, the growth illustrated by aggregated data did not match up to any one individual. When Larsen-Freeman looked at individual learners' data, she found that they actually exhibited diverging patterns of development, perhaps due to different allocations of attentional resources (p. 601). Despite a high degree of intra-individual variability, Larsen-Freeman isolated two "preferred paths" of development: some learners focused on lexical variety, while the others emphasized and improved most on grammatical complexity, to differing extents. Overall, all five learners had differential success, with each person following a somewhat different growth trajectory. Her rate of change calculations on the CAF measures revealed that the rate of change fluctuates for different learners at different times (p. 604), consistent with DST's predictions of non-linear change.

The qualitative results of Larsen-Freeman's study are also conducive to a CAS characterization of an L2. Although space limits my review, each of the learners exhibited a high degree of intra-individual variation between performances, with differences in variegated categories such as "morphemes, words, phrases, clauses, partial utterances, abstract semantic categories, etc." (p. 608). Since learners are performing the same task multiple times, variation

between performances may indicate “bifurcations,” signaling instability that can precede a phase shift in the CAS (p. 611). Larsen-Freeman suggests that such moments of variation are optimal times for pedagogical intervention. Although she did not run any correlations on the qualitative data, her article reminds researchers of the necessity to adopt both macro- and micro-level perspectives to capture development in a CAS such as an L2. Macro-level cross-sectional research illustrates “the grand sweep of development” with global trends and similarities across learners, while micro-level longitudinal case studies give a bottom-up view, including all the “messy little details” and intra-individual differences (p. 613), which Larsen-Freeman argues are all significant in a complex, dynamic system. In her words, “The messiness is not ‘noise’, but rather a natural part of dynamically emergent behavior assembled by the individual with a dynamic history of engaging in such tasks, with his or her self-identified (or jointly-identified) target of opportunities for growth” (p. 615). With respect to the CAF constructs, no particular subsystem of language has a priori precedence over any other. Instead, they interact in variegated ways, dynamically adapting to the environment and demands of a given moment and context. As Thelen and Bates (2003) summarize, developmental change seems “not so much the stage-like progression of new accomplishments as the waxing and waning of patterns, some stable and adaptive and others fleeting and seen only under special conditions” (p. 380).

2.5.5 Individual differences

Despite their common L1, comparable initial L2 proficiency, and similar instructional environments, the learners studied by Larsen-Freeman (2006) exhibited different paths of interlanguage development. In DST, there can be no single variable that can “cause” learners’ dynamic L2 systems to undergo a phase shift in one direction or another, as every single aspect

of the system is interconnected. At the same time, even if learners' L2 CAS are all unique, it is still possible that "preferred paths" of development may manifest themselves in different types of learners if enough learners are considered on a micro-level of analysis.

Skehan (1998) distinguishes between three types of learners who vary with respect to attention paid to form and meaning. One group of learners balances attention to meaning and form throughout development. They are able to

> [S]witch attention judiciously so that their interlanguage system is more likely to be regularly reviewed (leading to a more 'open' and permeable system in general) but attention is also devoted to integrating language, on an exemplar-base, so that natural communication is achieved. (p. 269)

But learners in the other groups do not exhibit such balanced progress because their processing preference (analytic vs. memory orientation) is associated with a prioritization of formal goals over communication goals and vice versa. The analytic learners focus on form over meaning/communication and tend to be excessively rule-oriented. Although they may achieve complexity in their knowledge representations, they may not be able to implement it in online oral production. In contrast, the memory-oriented learners focus on meaning over form and "may have acquired communicative fluency too early, with the result that fossilization makes later progress difficult… as a result of strategic competence and lexicalized communication becoming too effective" (p. 270). Although Skehan undoubtedly oversimplifies in a model of interlanguage development that may not apply to all learners, his parsimonious account of learner differences may help in understanding one cause for the high degree of individual variation found in CAF research.

In addition to meaning vs. form orientations, other individual differences such as affective factors, age, gender, initial proficiency, motivation, language background, time spent in a TL country, and communicative orientation may all contribute to different performances and

development over time. In the current research, I am most concerned with language background (L1), because L2 learners engage in SLA with firmly entrenched L1 patterns (N. Ellis, 2006). Neural commitment to these patterns results in cross-linguistic influence that manifests itself in terms of the paths of traversing developmental sequences, relexification, overgeneralization, avoidance, overproduction, and hypercorrection (Odlin, 1989). L1 can also tune learners' perceptual mechanisms so that their learned attention blocks them from perceiving relevant differences in the L2 that are irrelevant in the L1 (i.e., /l/ vs. /r/ distinction for Japanese learners of English; cf. N. Ellis, 2006).

Despite the large degree of individual variation observed by Larsen-Freeman (2006) even among learners with a common language background, it is likely that there is also potential for L1-based differences. This is particularly relevant for this study, as I am comparing L1 Arabic and L1 Chinese learners, who have contrasting communicative orientations and learning styles, likely due to their divergent cultural backgrounds. These differing learner orientations are discussed in Section 4.3.3.

3.0 MORPHEME STUDIES

Beginning in the 1970s, a number of studies were carried out to investigate the order of acquisition of grammatical functors in first language acquisition of English including articles, copulas, nominal marking like plural *–s*, and verbal inflection such as progressive *–ing*, regular past *–ed*, and third person singular present *–s*, among others (Brown, 1973; de Villiers & de Villiers, 1973). These so-called "morpheme studies" have since become benchmarks in the fields of both first and second language acquisition, though many applied linguists continue to debate the significance of and agreement between such studies (Cook, 1993; R. Ellis, 2008; Gass & Selinker, 2008; Long & Sato, 1984).

3.1 FIRST LANGUAGE ACQUISITION

In his seminal book "A First Language," Roger Brown (1973) studied the longitudinal development of three children, Adam, Eve, and Sarah, as they acquired English as their L1. He collected naturally occurring oral output every one to two weeks starting at age two. Brown examined the learners' linguistic development through two lenses: mean length of utterance (MLU) and a morphemic analysis. MLU is a measure of the average number of morphemes per utterance, which Brown considered "an excellent simple index of grammatical development because almost every new kind of knowledge increases length" (p. 53). Brown found that the

children's MLU value increased consistently with chronological age, although Eve experienced growth in MLU at a slightly earlier age than Adam and Sarah, which Brown credited to individual differences.

The second measure of development investigated by Brown (1973) was a morphemic analysis that considered suppliance of grammatical functors in obligatory contexts. He referred to such criteria of acquisition as a type of "output-where-required." Brown justified this measure of suppliance in obligatory contexts as follows:

> Each obligatory context can be regarded as a kind of test item which the child passes by supplying the required morpheme or fails by supplying none or one that is not correct. This performance measure, the percentage of morphemes supplied in obligatory contests, should not be dependent on the topic of conversation or the character of the interaction. (p. 255)

For each grammatical functor, Brown calculated an accuracy score in order to operationalize the criterion of whether a TL feature had been acquired. According to his rigorous criteria, a given TL feature was only acquired if a learner supplied it 90% of the time on three consecutive data collection points. After collecting all data, Brown ranked the grammatical functors in order of acquisition for each child and then calculated Spearman rank-order correlations among the three children. Brown found a high level of consistency among the orders of acquisition, with Spearman rho values as follows: for Adam and Sarah, r_s = .88; for Adam and Eve, r_s = 0.86; and for Eve and Sarah, r_s = 0.87.

Given all three children's data, Brown found that grammatical functors were acquired in a predictable order of present progressive *–ing*; *in/on*; plural *–s*; past irregular; possessive 's; uncontractible copulas (*is, am, are*); articles (*a, the*); past regular *–ed*; third person sing. *–s*; third person irregulars (*does, has*); uncontractible auxiliary; contractible copula; and finally, contractible auxiliary (p. 271). However, such an observation is only as significant as the

author's explanation for it. Brown could not directly attribute the order to the frequency of grammatical functors in parental speech because the three sets of parents exhibited similar frequencies to each other, but these did not correspond to the order of acquisition, with a Spearman's rho correlation of only $r_s = 0.26$ (pp. 358-359). Brown therefore attributed the order of acquisition to the interaction of grammatical complexity and semantic complexity of these various TL features.

In the same year, De Villiers and De Villiers (1973) attempted to isolate an acquisition order in a cross-sectional study of 21 children (aged between 16 and 40 months) acquiring English as their L1 by analyzing speech samples from two 1.5-hour play sessions. Instead of using Brown's (1973) stringent 90% suppliance as the criterion, they simply ranked the grammatical functors according to relative accuracy. De Villiers and De Villiers' participants exhibited a morpheme order similar to Brown's, with the magnitude of the correlation between their learners comparable to the correlations that Brown found between Adam, Eve, and Sarah. De Villiers and De Villiers discovered that MLU was a better predictor of morpheme acquisition than chronological age, but like Brown (1973), they could not pinpoint a single cause for the morpheme acquisition order. De Villiers and De Villiers concluded, "[T]he order of acquisition may best be predicted by some combination of grammatical and semantic complexity, frequency, and perceptibility in speech," (p. 277), with no one factor claiming primary importance in determining the acquisition of the morphemes.

3.2 CHILD SECOND LANGUAGE MORPHEME STUDIES

Dulay and Burt (1973) carried out a similar cross-sectional study of 151 L1 Spanish speakers aged six to eight who were acquiring English as an L2 while living in America. The learners comprised three separate groups: Chicano children studying in Sacramento, CA; Mexican children living in Tijuana but attending school in San Ysidro, CA; and Puerto Rican children in New York City. Dulay and Burt collected oral output data using the Bilingual Syntax Measure (BSM), a structured conversation technique originally designed to measure L2 proficiency. When the BSM is administered, learners describe seven colorful cartoon pictures and answer 33 accompanying questions, designed to elicit certain structures in obligatory contexts. Dulay and Burt measured accuracy of a given grammatical functor (e.g., progressive *–ing*) according to suppliance in obligatory contexts (SOC), with points assigned according to the schema illustrated in Table 3 (p. 254):

Table 3. Dulay and Burt's (1973) suppliance in obligatory contexts (SOC) scoring schema

Suppliance	Score	Example
No functor supplied	0	she's dance__
Misformed functor supplied	0.5	she's dance*s*
Correct functor supplied	1.0	she's danc*ing*

A total SOC score for each grammatical functor was calculated as follows:

$$(8)\ SOC_1 = \frac{1 * (n \text{ correct suppliance in obligatory context}) + 0.5 * (n \text{ misformations in obligatory context})}{1 * (n \text{ total obligatory contexts})}$$

The resulting value is then multiplied by 100 to yield an accuracy percentage score for each grammatical functor.

Dulay and Burt (1973) found that despite different backgrounds and environments, the three groups of L1 Spanish learners exhibited a similar accuracy order on the following eight

structures: plural *–s*; progressive *–ing*; copula *is*; articles *a*, *the*; auxiliary *is*; irregular past; third person singular *–s*; and finally, possessive *–s*. What is significant is how this L2 order differs from the L1 English order found by Brown (1973) and De Villiers and De Villiers (1973). Dulay and Burt (1973) attributed the different order to the fact that "the older L2 learner need not struggle with the same kinds of semantic notions already acquired in earlier childhood" (p. 252). However, the authors claimed that beyond semantics, child second language acquisition was similar to child first language acquisition, a claim that came to be known as the L1 = L2 hypothesis. Dulay and Burt (1974) attribute the accuracy order to *creative construction*, which they define as

> [T]he process in which children gradually reconstruct rules for speech they hear, guided by universal innate mechanisms which cause them to formulate certain types of hypotheses about the language system being acquired, until the mismatch between what they are exposed to and what they produce is resolved. (p. 37)

In order to illustrate that the child L2 morpheme order was a result of universals at play and not the language background of the learners, Dulay and Burt (1974) extended their study to include 55 L1 Cantonese children studying in Chinatown, NY and 60 L1 Spanish children studying in Long Island, NY, all aged six to eight. Dulay and Burt relied on the same SOC scoring technique, but this time modified the point system as follows:

$$(9)\ SOC_2 = \frac{2 * (n \text{ correct suppliance in obligatory context}) + 1 * (n \text{ misformations in obligatory context})}{2 * (n \text{ total obligatory contexts})}$$

They calculated both group score and group mean figures for each functor. Group scores give a total average accuracy score for each morpheme in each of the two L1 groups, including all contexts from all learners (pp. 44-45). Group means also give a average accuracy score for each

morpheme by L1 group, but a group mean is calculated using data only from the learners who produced (or omitted) a given functor on three or more occasions, thus eliminating from the sample size learners who only produced or omitted something one or two times (pp. 45-46).

Although the L1 Spanish speakers were overall slightly more accurate on all grammatical functors by about 10%, Dulay and Burt (1974) found significant correlations between the two language background groups' rank orders regardless of the measure. Relying on the group score method, they found that Spanish and Chinese rank orders had a Spearman rho correlation of .95, while the group means method rank orders of the two L1 groups had a correlation of .96, both at $p < .001$ (p. 50). The accuracy order for both groups of learners was pronoun case; articles *a, the*; progressive *–ing*; contractible copula *'s*; short plural *–s*; contractible auxiliary *'s*; past regular *-ed*; past irregular; long plural *–es*; possessive *'s*; and finally, third person singular present *-s*. In accounting for the order, Dulay and Burt (1974) propose "universal strategies" that are "sufficiently abstract and comprehensive so as to predict acquisition orders based on different types of language input, such as languages other than English, or types of speech exposure other than natural speech" (p. 52). In other words, they claim that the order is predictable and is the result of L2 learning strategies common to all children acquiring English in natural host country environments, regardless of and minimally influenced by their L1. In the years that followed, numerous other morpheme studies were carried out on children acquiring English in a variety of settings and with different L1s (Hakuta, 1974, 1976; Kessler & Idar, 1979; Kjaarsgard, 1979; Mace-Matlock, 1977, 1979; Riddle, 1993; Rosansky, 1976, among others). Despite the participants' different language backgrounds, most of these studies (excluding Hakuta, 1976) found that children acquiring English as an L2 acquire grammatical morphemes in what Dulay and Burt (1974) called a "universal order" (p. 50).

3.3 ADULT SECOND LANGUAGE MORPHEME STUDIES

The next reasonable step was to extend the morpheme order studies to adults learning English as an L2. Bailey, Madden, and Krashen (1974) carried out a study similar to Dulay and Burt (1974) using the BSM on adult learners of English, with one group of 33 native speakers of Spanish, and one group of 40 speakers with eleven "non-Spanish" L1s including Greek, Persian, Italian, Turkish, Japanese, Chinese, Thai, Afghan, Hebrew, Arabic, and Vietnamese. Despite the imperfect design that conflated many speakers with unrelated L1s, the researchers found that morphemes still followed what Krashen (1977) later called a "natural order" of accuracy of progressive *–ing*; contractible copula *'s*; plural *–s*; articles *a*, *the*; contractible auxiliary *'s*; past irregular; third person singular present tense *-s*; then possessive *'s*—all with minimal L1 influence.

Numerous other studies were carried out on adult English language learners of diverse L1s, and nearly all found similar orders (Andersen, 1976, 1977; Ball, 1996; Brown, 1983; Fathman, 1975; Fuller, 1978; Houck, Robertson, & Krashen, 1978; Krashen, Butler, Birnbaum, & Robertson, 1978; Krashen, Houck, Giunchi, Bode, Birnbaum, & Strei, 1977; Krashen, Sferlazza, Feldman, & Fathman, 1978; Larsen-Freeman, 1975; Lightbown, 1983; Lightbown, Spada, & Wallace, 1980; Makino, 1979; Pak, 1987; Pica, 1983; Rosaldo, 1986; and Rosansky, 1976, among others). This phenomenon of a natural order demanded an explanation. Based on longitudinal orders of emergence and cross-sectional accuracy orders found in the earlier research, Krashen (1977) postulated a natural order of four stages common to adult learners, illustrated in Figure 1:

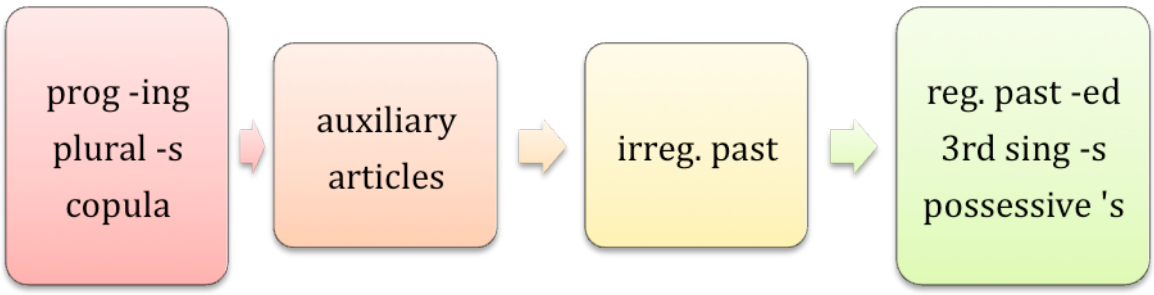

Figure 1. Krashen's (1977) natural order for adult ESL learners

Krashen (1977) observed the "amazing amount of uniformity across all studies" (p. 148) but failed to give an explanation for this common order. Nonetheless, research on the natural order of emergence continued in the decades that followed, but was not immune to criticism regarding both the significance and methodological soundness of such studies, discussed below.

3.4 THEORETICAL ISSUES WITH THE MORPHEME STUDIES

Since the emergence of the morpheme studies, applied linguistics have criticized this line of research for a number of reasons (Cook, 1983; Long & Sato, 1984; see also Gass & Selinker, 2008; Larsen-Freeman & Long, 1991).

Porter (1977) was among the first to argue that the observed order was a test artifact of the BSM. However, this claim was refuted by other studies that used different types of elicitation but still obtained the same order. Other elicitation methods include the Second Language Oral Production English (SLOPE) test (Fathman, 1975; Krashen, Sferlazza, Feldman, & Fathman, 1978); the MAT-SEA-CAL Oral Proficiency Test (Mace-Matluck, 1977); free composition data (Houck, Robertson, & Krashen, 1978; Krashen, Butler, Birnbaum, & Robertson, 1978); sentence repetition tasks, listening comprehension tests, a reading cloze passage, and a writing test (Larsen-Freeman, 1975); among others.

Another common critique is the linguistic heterogeneity of the grammatical functors and structures under investigation (Krashen, 1977). They consist of both bound and free NP and VP morphemes and many researchers neglect to analyze morphemes in subsets, which may reveal more about developmental patterns (Andersen, 1978; Brown, 1983; Zobl & Liceras, 1994). In addition, most of the morpheme studies group *a/an* and *the* together in the ARTICLE category, neglecting Ø, but Andersen (1977) showed that diverse English language learners (ELLs) behaved differently with *a* and *the*, so the articles should not have been grouped together as a single grammatical functor in the first place. Furthermore, the morphemes under investigation are obviously specific to English, so they preclude any possible cross-linguistic generalizations. If anything, these studies reveal something about ESL, not SLA (Larsen-Freeman & Long, 1991). Moreover, what they reveal is quite limited, as the elements of English grammar that are studied constitute a small part of the language as a whole.

The design of the morpheme studies, specifically the scoring method of SOC, has also come under fire. SOC scoring compares learner language to TL norms, thus committing the Comparative Fallacy (Bley-Vroman, 1983). This is an issue because often learners' IL is systematic in its own rite. As Goldschneider and DeKeyser (2001) explain,

> Measuring the accurate use of grammatical functors by an ESL student by comparing it to an ideal of the target language risks denying the internal logic of the student's interlanguage. What is "accurate" in the target language may have nothing to do with what is accurate in the student's own grammar at that point in time. (p. 17)

Therefore, SOC measures may do little to illuminate the internal structure of a learner's interlanguage and its development over time.

Similarly, other critics have questioned whether it is theoretically sound to use SOC-based accuracy scores as a measure of acquisition because such a measure disregards incorrect

overgeneralization of given structures. Overuse and overgeneralization are significant because acquisition of a morpheme can be defined not only as knowing when to use the morpheme, but also when not to use it. For this reason, Pica (1983) used another scoring method called Target-Like Use (TLU), which is calculated as follows (p. 474):

$$(10)\ \text{TLU} = \frac{n \text{ correct suppliance in obligatory contexts}}{n \text{ obligatory contexts} + n \text{ suppliance in nonobligatory contexts}}$$

Although TLU still compares learner language to target language norms, it addresses the problem of overgeneralization, as the overuse of a morpheme in a non-obligatory context still lowers the overall accuracy score. Still, the high degree of correlation between orders obtained using SOC and TLU measures helps counter the criticism against morpheme studies that neglected to consider overuse (R. Ellis, 2008; Larsen-Freeman & Long, 1991). However, the correlation between SOC and TLU does not hold if articles are considered separately. For example, Lu (2001) suggested that for Chinese learners of English, TLU is a more reliable acquisition measure. TLU analyses reveal the order of *the* emerging before *a,* which in turn emerges before null article Ø. In contrast, SOC serves better as an index of accuracy, where *the* and *a* have similarly higher accuracy than Ø (p. 43).

Another critique concerns the theoretical grounding of the morpheme studies and the assumed correspondence between accuracy order (obtained in cross-sectional studies) and order of acquisition (i.e., developmental sequence, as found in longitudinal studies). Rosansky (1976) performed a 10-month longitudinal study on six L1 Spanish learners of English and found that their order of acquisition of morphemes over time coincided neither with their cross-sectional (grouped) accuracy order, nor with the natural order reported for other groups of learners (Bailey et al., 1974; Dulay & Burt, 1973, 1974; Krashen, 1977). Similarly, Hakuta (1976) collected

longitudinal data from a five-year-old Japanese learner of English, but the acquisition order observed deviated significantly from that found by Dulay and Burt (1973, 1974) in their cross-sectional studies. Krashen (1977) addressed these longitudinal vs. cross-sectional deviations by referring to Hakuta's learner as "idiosyncratic" (see Dulay, Burt, & Krashen (1982) p. 207) and pinpointing methodological problems in Rosansky's study. More specifically, Krashen (1977) argued that the natural order only holds in studies with ten or more obligatory contexts per morpheme. Calculations performed with fewer than ten obligatory contexts (as in Rosansky, 1976) can produce unreliable results, especially given the high degree of individual variability (cf. Larsen-Freeman & Long, 1991, p. 90).

The final criticisms of the morpheme studies challenge their statistical significance. J. D. Brown (1983) notes that inferential statistics employed in determining rank orders (Kendall or Spearman rank order correlations) are a "weak" type of inferential test. Even if two orders were statistically significantly related, they could still differ in significant ways – which they did. What's more, many morpheme studies' authors only list the rank orders of the functors, without reporting the accuracy percentage on which the rank orders are based (e.g., Dulay & Burt 1973, 1974; Larsen-Freeman, 1975; Hakuta, 1976; Pak 1987). Goldschneider and DeKeyser (2001) warn that such rank orders can be misleading as they fail to indicate the relative distance between the percentages for each functor. For example, a morpheme that is one percent lower in accuracy than another morpheme is given a different ranking, just as a morpheme that is 30 percent lower (R. Ellis, 2008, p. 83). Thus, the notion of ranking may be misleading if one does not closely examine the numerical differences in accuracy between different functors.

The last statistical criticism is of the very design of morpheme studies themselves because they combine individuals' data together. Grouping data together based on L1,

proficiency level, instructional setting, etc. risks hiding a large degree of potentially significant variability at the individual level. In order to counter this claim, Andersen (1978) showed that individual and group morpheme data correlate significantly and proposed an implicational model for scoring grammatical functor data to capture that variability. However, not all studies adopt this implicational model and instead only consider group data. Far too often, meaningful individual variability is obscured by group data, leaving researchers with a much simpler, homogenous picture of acquisition than is truly the case. For this reason, this MA thesis considers output at both the group and individual level.

3.5 THE SEARCH FOR AN EXPLANATION

Despite the theoretical and methodological issues with the morpheme order, Larsen-Freeman and Long (1991) have concluded

> [T]he morpheme studies provide strong evidence that ILs [interlanguages] exhibit common accuracy/acquisition orders. Contrary to what some critics have alleged, there are in our view too many studies conducted with significant methodological rigor and showing sufficiently consistent general findings for the commonalities to be ignored. As the hunter put it, “There is something moving in the bushes.” (p. 92)

The question, then, is what. Larsen-Freeman (1975, 1976) suggested input frequency as an explanatory factor, but noted the insufficiency of frequency effects, as articles are the most frequent functor to which learners are exposed, but they are not the first forms mastered. Larsen-Freeman (1975) thus deduced that a “single explanation seems insufficient to account for the findings” (p. 419).

Some researchers have sought to explain the order through combinations of characteristics of the grammatical functors, including semantic and syntactic complexity, perceptual saliency, functional transparency, and input frequency (Goldschneider & DeKeyser, 2001; Zobl & Liceras, 1994; see also Larsen-Freeman, 1976; Lightbown, 1983; Long & Sato, 1983). Other applied linguists have suggested that the order of acquisition of given L2 structures is an epiphenomenon observed as a result of underlying universal processing constraints (Pienemann & Johnston, 1985). Larsen-Freeman and Long (1991) note, "The universality of the constraints potentially explains the commonalities across learners of both the morpheme accuracy/acquisition order and developmental sequences" (p. 91). However, a more thorough analysis of processing constraints is beyond the scope of this thesis.

3.5.1 Functional categories

The different L1/L2 English morpheme order explanations demanded an explanation. Krashen, Bailey, and Madden (1975) noted that auxiliary and copula ranked higher in L2 than L1 orders and suggested enhanced memory capacity and prior linguistic knowledge in L2 learners as possible reasons for the different rankings. This explanation assumes that functional categories (e.g., complementizers, prepositions, determiners, etc., in contrast to lexical categories like noun, verb, adjective, adverb) are available to the learner thanks to prior linguistic experience (i.e., mastery of an L1), but Krashen et al. failed to provide a data-driven explanation. Furthermore, they denied any importance to the instantiation of functional categories in a learner's L1 and the potential for interference depending on how the L1 and L2 match up.

In a study comparing accurate L2 production of past verbal forms, Rod Ellis (1987) noted that regular past verbs are formed via the application of a syntactic rule, while irregular forms are

stored lexically. Ellis found that the relative accuracy on regular vs. irregular past forms was differentially affected by task conditions such as planning. More specifically, the regular past was significantly affected by task conditions, with a great decline in accuracy as planning time decreased, while irregular past performance hardly declines at all. Ellis argued that these results could also be the manifestation of U-shaped learning, where progress is marked by a step backwards and necessary restructuring.

Zobl and Liceras (1994) were among the first to propose a systematic explanation for the L2 English morpheme order and its divergence from the L1 English order by appealing to syntactic category (lexical vs. functional) and bound morphemes vs. free functors. The authors account for the different L1/L2 orders by arguing that functional categories in the L1 emerge in categories over time, with nominal categories preceding verbal categories. In contrast, L2 learning involves cross-category development of both nominal and verbal functional categories and the affixes where they are realized, which is consistent with the hypothesis that functional projections are available from the start (Poeppel & Wexler, 1993). Zobl and Liceras also distinguish between functional and lexical morphemes in terms of how they implement functional categories. In L1, functional and lexical morphemes "play a coequal role" (p. 162), while in L2, functional projections are already available to the learner, so inflectional morphemes need not mark functional categories. Finally, the authors observe that free morphemes emerge before affixes in L2 development, suggesting that when affixes finally *do* emerge, their movement "seems to play a key role in the development of those affixes having a syntactic function" (p. 162). To summarize, the combination of these factors produces an L2 order of lexical items being acquired before functional items, and within each group, free morphemes acquired before bound ones.

In terms of implications, Zobl and Liceras' (1994) analysis suggests access to Universal Grammar (UG) despite how learners' L1 may differ from the L2 with respect to parameterization and how certain features bundle together (see Lardiere, 2009). The authors conclude with a claim that L2 learners of English have access to functional categories from the start and need not learn them as such, but only their language-specific realization in English. However, these syntactic and semantic factors are only part of the explanation and prove unsatisfactory. Therefore, it is necessary to consider other factors such as phonological salience, morphological regularity, input frequency, redundancy,[6] and last but not least, L1.

3.5.2 Universals

In order to account for the natural order observed in numerous morpheme studies, Goldschneider and DeKeyser (2001) performed a meta-analysis of 12 studies containing 924 subjects with 28 different L1s, selected from a pool of 25 ESL studies carried out between 1973 and 1996. Their criteria restricted their selection to studies that involved only oral production data gathered in ESL settings, and, for obvious reasons, studies that report SOC percentages for each functor and not just rank orders. In their meta-analysis, the authors considered six functors common to these 12 studies: present progressing *–ing*; plural *–s*; possessive 's; articles *a*, *an*, *the*; third singular present *–s*; and regular past *–ed*. They attempt to account for the observed orders by identifying

[6] Here, redundancy refers to whether a given grammatical functor is necessary to communicate meaning (e.g., plurality in *I love my brother**s***) or is redundant, as the referential information is included elsewhere (e.g., by the numeral in *I love my **three** brothers*). The functional hypothesis would predict more suppliance in necessary than redundant contents. Past research has indicated that redundant marking generally increases with proficiency (Young, 1993). However, Schepps (2013) observed that context may play opposing roles for different groups of learners, and that individual variability is an important factor in interlanguage morphology.

and quantifying five broad predictors: perceptual (i.e., phonological) salience, semantic complexity, morphophonological regularity, syntactic category, and frequency.

Perceptual salience refers to how easy it is to perceive or hear a given structure, which Goldschneider and DeKeyser (2001) operationalized by number of phones, syllabicity, and sonority. They assumed that the more perceptually salient a functor is, the easier it is to acquire. Next, semantic complexity refers to how many meanings are expressed by a particular form; for instance, plural *–s* expresses number, while third singular present *–s* expresses person, number, and tense. Morphophonological regularity refers to the extent to which functors are affected by their phonological environment, assuming that more phonologically regular functors are easier to acquire. It is operationalized through the number of phonological alternations and whether a functor is homophonous with others. Syntactic category is similar to Brown's (1973) notion of syntactic complexity and refers to the characteristics of each functor from a Functional Category theory perspective (Zobl & Liceras, 1994). Finally, input frequency refers to the number of times a functor occurs in speech addressed to the learners. Because it was impossible to quantify and analyze all of the input to which the 924 learners in these 12 studies were exposed, Goldschneider and DeKeyser instead used Brown's (1973, p. 358) pooled frequency data from the three sets of parents. These determinants, their operationalization, and the accompanying predictions are summarized in Table 4.

Table 4. Goldschneider and DeKeyser's (2001) multiple determinants

Determinant	Operationalization	Predictions
Perceptual salience	Number of phones	More phones in a functor → more perceptually salient → easier to acquire
Perceptual salience	Syllabicity	Functors containing a vowel in the surface form → more perceptually salient → easier to acquire
Perceptual salience	Sonority	Functors that are more sonorous → more perceptually salient → easier to acquire
Semantic	Number of meanings	Forms with more meanings → harder to

complexity		acquire
Morphophonological regularity	Number of phonological alternations	More alternations → less phonologically regular → harder to acquire
Morphophonological regularity	Homophony with other functors	If homophonous with other functors → harder to acquire
Syntactic category	Lexical vs. functional Bound vs. Free (Zobl & Liceras, 1994)	Lexical, free items → easiest to acquire Lexical, bound items → easy to acquire Functional, free items → hard to acquire Functional, bound items → hardest to acquire
Input frequency	Brown's (1973) parental speech corpus	Functors that are more frequent in the input → easier to acquire

Performing a multiple regression analysis of 924 subjects' orders, Goldschneider and DeKeyser found that the interaction of these predictors (all but frequency & morphophonological regularity and perceptual salience & semantic complexity) had a significant intercorrelation ($p < .05$). In terms of explanatory power, the authors found that 71% ($R = .84$, $R^2 = .71$, $p < .001$) of the variance observed across all learners could be explained simply by the effects of these five main determinants. In their discussion, Goldschneider and DeKeyser (2001) argue that the factors themselves "all constitute aspects of salience in a broad sense of the word" (p. 35), an intuitive fact that long demanded statistical backing. To date, Goldschneider and DeKeyser's study is likely the most comprehensive explanation for the observed natural order and among the best retorts to morpheme study critiques by proposing salience as the "ultimate predictor of the order of acquisition" (p. 36).

Because of its nature as a meta-analysis of 924 learners' data from 28 typologically diverse L1s, the authors were unable to systematically investigate L1 influence (also known in the literature as L1 transfer or interference, here used interchangeably). Goldschneider and DeKeyser (2001) noted that the number of diverse L1s represented in their pooled data greatly reduces the possibility of skewed results due to L1 influence (p. 31), the potential predictor to which I now turn in Section 3.5.3. However, Luk and Shirai (2009) observed that Goldshneider

and DeKeyser's meta-analysis of the natural order may have actually been skewed by L1 because 354 of their 924 participants had L1 Spanish and were therefore over-represented in the pool (p. 739).

3.5.3 The role of the first language

The role of the L1 in determining acquisition or accuracy orders has been the subject of considerable disagreement since the publication of the first ESL morpheme studies. Dulay and Burt (1973) reported that only 3% of the errors made by the children in their study could be attributed to L1 Spanish interference. In contrast, Larsen-Freeman (1975) found that Japanese-speaking learners of English had lower scores on articles than other learners, which she argued is because Japanese is a language without an article system. Similarly, Hakuta's (1976) longitudinal study on an L1 Japanese girl learning English revealed an order of acquisition that deviated from Dulay and Burt's (1973, 1974) findings, with articles acquired especially late because of the learner's difficulty with the definite/indefinite contrast, which does not exist in her L1. Hakuta and Cancino (1977) later proposed that in general, an L2 learner whose L1 does not make the same semantic distinctions as the L2 with regard to particular morphemes or functors will have a more difficult time acquiring such morphemes.

Other research supported such a proposal. For example, Andersen (1977) attributed a large degree of L1 influence to Spanish speakers' acquisition of articles and possessive *'s* and later concluded (1978) that L1 influence is clearly "a factor that must be taken into consideration as one of the factors that could interact with morpheme acquisition and accuracy orders" (p. 267). Likewise, Pak (1987) found that a group of Korean-speaking children learning English in Texas had much greater difficulty with the definite article and plural *–s* than the L1 Chinese and

L1 Spanish children studied by Dulay and Burt (1973, 1974). Shin and Milroy (1999) also found that Korean-speaking children's English acquisition order differed from L1 Chinese and Spanish learners' order, but was similar to the Japanese order found by Hakuta (1976). In particular, the Korean children did well on pronoun case and possessive *'s*, but performed poorly on plural *–s*, articles, and third singular present *–s*. This led Shin and Milroy to conclude that there are L1 specific influences on SLA. As N. C. Ellis (2006) summarizes, "The fact that Japanese and Korean are morphosyntactically very similar confirms these language specific influences on L2 acquisition: L2 acquisition is clearly affected by the transfer of learners' knowledge of their first language" (p. 187). All of these deviations from the natural order suggest that the acquisition order of morphemes cannot be explained entirely by universals, and the relationship between L1 and L2 is more nuanced than previously thought.

More recent commentaries (R. Ellis, 2008; Gass & Selinker, 2008) have suggested that the traditional tendency to disregard L1 in accounting for ESL morpheme orders (c.f. Dulay and Burt 1973, 1974; Krashen, 1977; Dulay, Burt, & Krashen, 1982) may have been a product of the times. During the 1970s, there was a change in thought underway among both theoretical and applied linguists. Many of those influenced by Chomsky adopted a universal stance and moved away from behaviorist accounts of SLA. This universalist perspective was adopted by many L2 researchers such as Dulay, Burt, and Krashen, who sought to explain the observed orders through innate, universal strategies they called creative construction. Because L1 transfer was ill-defined and strongly associated (if not equated) with behaviorist theory and contrastive analysis at the time, many researchers effectively "threw the baby out with the bathwater" as they disregarded both behaviorism and any potential influence from the L1 in favor of an innate, universal perspective (Gass & Selinker, 2008, p. 135). At the same time, UG-based frameworks have

included a role for the L1 (White, 1989), including a full transfer, full access approach (Schwartz & Sprouse, 1996), while cognitive views of SLA also incorporate a significant and systematic role for the L1 (Kellerman & Sharwood-Smith, 1986), more recently from a neural, connectionist viewpoint (cf. N. Ellis, 2006).

Luk and Shirai (2009) performed an analysis of data from previous morpheme studies, first selecting data from the original pool of 25 studies by Goldschneider and Dekeyser (2001), and then adding other studies that included data for more than 8 morphemes in Krashen's natural order from L1 Japanese, Korean, Chinese, and Spanish learners of English to test the effect of L1 in morpheme acquisition. The authors compared rank orders from L1 Japanese (Hakuta, 1976; Izumi & Isahara, 2004; Koike, 1983; Makino, 1979; Nuibe, 1986; Sasaki, 1987; Shirahata, 1988), L1 Korean (Pak, 1987; Shin & Milroy, 1999), and L1 Chinese (Dulay & Burt, 1974; Mace-Matluck, 1979) learners of English to those of L1 Spanish learners (Andersen, 1978; Bailey et al., 1974; Dulay & Burt, 1973, 1974; Mace-Matluck, 1979; Pica, 1983; Rosansky, 1976). They found a general trend that Japanese, Korean, and Chinese learners of English acquire possessive *–s* earlier, and articles and plural *–s* later than Spanish learners and the natural order. They attribute the different orders to L1 influence, because the three Asiatic languages have possessive structures similar to English, but lack comparable definite/indefinite articles and plural morphemes, while Spanish has the same definite/indefinite article distinction and even a homophonous plural morpheme, but lacks a possessive construction analogous to English. Luk and Shirai's (2009) study was influential because it was the first to provide evidence that suggests "L1 transfer is much stronger than is portrayed in many SLA textbooks and that the role of L1 in morpheme acquisition must be reconsidered" (p. 721).

Luk and Shirai's (2009) findings of "notable L1 effects" are an important contribution to the field and the move away from a universal morpheme acquisition order impervious to L1 influence (p. 738). They adopt a cognitive view of L1 transfer in which once an L1 is acquired, learners cannot process an L2 without the filter of the L1 (p. 740). More specifically, Luk and Shirai (2009) explain that:

> [I]n various linguistic domains, learning a native language involves acquiring the ability to process it efficiently and learning to ignore—or losing the ability to make—the distinctions that are unimportant in the language.… because L2 learning occurs through the filter of the L1 network, it is only natural that there are very different acquisition orders for different L1 groups, rather than a universal natural order. (p. 742)

They echo N. C. Ellis (2006), who illustrated how such L1 network filter effects can account for why certain L2 forms fail to become intake in a learner's L2 processing "because of one of associative learning factors of contingency, cue competition, or salience, or because of associative attentional tuning involving interference, overshadowing and blocking, or perceptual learning, all shaped by the L1" (p. 165). In any case, both Luk and Shirai's and Ellis' articles indicate a significant role for the L1 and that any supposed universal learning mechanisms must consider L1 influence. Consequently, in Section 4.3.2, I discuss potential correspondences in Arabic and Chinese for the six grammatical functors under investigation in the current study.

4.0 THE STUDY

4.1 METHODOLOGY

This research considers the spontaneous oral output of 30 learners of English enrolled in the University of Pittsburgh English Language Institute (ELI), described in 4.2, as they progress from a low to high intermediate level of English proficiency. The study involved the coding and quantitative analysis of six two-minute semi-spontaneous speeches per learner, with observations evenly spaced over two semesters of study (an eight-month period). The speeches comprise part of the Recorded Speaking Activity (RSA), an obligatory part of the speaking curriculum, described in 4.4.1. All speeches come from the ELI Online Database, a searchable corpus of written and spoken student data from 2006 to the present.

The participants' six speeches were first coded into AS-units and clauses, with errors as well as fluency breakdown and repair characteristics recorded in order to calculate CAF measures. Then, the data were coded for specific accuracy scores on six grammatical functors (plural *–s*; articles *a/an* and *the*; past regular *–ed*; irregular past; and third person sing. present *–s*) in terms of correct suppliance, misformed suppliance, omission in obligatory contexts, and oversuppliance in non-obligatory contexts. The morpheme data from the three speeches at Level 3 and the three speeches at Level 4 were collapsed in order to allow enough contexts to draw comparisons.

4.2 THE ENGLISH LANGUAGE INSTITUTE

The English Language Institute (ELI) is a CEA-accredited Intensive English Program (IEP) associated with the Linguistics Department at the University of Pittsburgh. Full-time ELI students are concurrently enrolled in reading, writing, speaking, listening, and grammar classes at either Level 3 (low intermediate), Level 4 (high intermediate), or Level 5 (low advanced), although some learners may be enrolled in different courses at two levels depending on their initial proficiency. Overall, the focus is on English for Academic Purposes (EAP), and this instructed SLA environment is comparable to other United States IEPs.

International students enroll at the ELI for diverse reasons. Some study English at the ELI in order to pass the TOEFL, IELTS, and/or GRE exams and begin undergraduate or graduate study at an American university, while others seek to improve their English language proficiency for professional or personal reasons. In any case, such ESL learners tend to be highly motivated (see Dörnyei, 2009) because they have opted to engage in non-obligatory study with at least 20 hours of lessons per week. While some learners are enrolled at the ELI for just one semester, others may study for four semesters, with the average enrollment of two semesters. Therefore, the data I consider is fairly representative of the typical student's enrollment span.

4.3 STUDENTS

The data in this study come from ELI students enrolled over at least two consecutive semesters from spring 2006 to spring 2010. Among the cohort of 30 learners, 15 have L1 Arabic and come from Saudi Arabia, while 15 have L1 Chinese and come from the People's Republic of China or

Taiwan. All of the Saudi participants were male, and 10 of the Chinese participants were female, giving a total gender breakdown of 66.67% male. The average age at data collection was 24.33, SD = 4.87, ranging from 18 to 37. More specifically, the Arabic learners' average age was 21.47, SD = 3.14, while the Chinese participants had a mean age of 27.20, SD = 4.66. Individuals' demographic information including age at data collection, gender, semesters of enrollment, and initial proficiency scores is included in Appendix A.

All learners began their studies at the ELI at Level 3 and continued at least through Level 4. The number of learners enrolled in each semester broken down by language background is summarized in Table 5.

Table 5. Number of learners per semester

Semesters of Enrollment	L1 Arabic Learners	L1 Chinese Learners	Total Number of Learners
Spring 2006, summer 2006	3	1	4
Summer 2006, fall 2006	8	2	10
Fall 2006, spring 2007	0	4	4
Spring 2007, summer 2007	1	0	1
Summer 2007, fall 2007	2	1	3
Fall 2007, spring 2008	1	2	3
Spring 2008, summer 2008	0	1*	1*
Summer 2008, fall 2008	0	2	2
Summer 2009, fall 2009	0	1	1
Fall 2009, spring 2010	0	1	1
Total participants (N) =	15	15	30

* Chinese learner 611 was omitted from the CAF analysis because his second Level 4 RSA was absent from the database.

Although the morpheme accuracy measures consider the output of all 30 learners, Chinese learner 611's second Level 4 speaking activity (from summer 2008) was absent from the database, likely because the student missed that assignment. Participant C 611 was therefore omitted from the pool for the CAF analyses, whose sample size was reduced to 29 learners, with six collection points each, yielding 174 total observations. In contrast, all 179 observations were

coded for the morpheme analysis. Each learner's specific topics and prompts are included in Appendix B.

4.3.1 Initial proficiency

Unlike past research that compared learners of varying initial proficiency levels (Spinner, 2011; Vercellotti, 2012) over the same semesters, the learners in this study were chosen from the ELI online database as they progressed from Level 3 to Level 4 because previous error analysis research (as well as anecdotal evidence) suggests that progress is often most marked as students advance from Level B1 to B2 in the Common European Framework (Thewissen, 2013), proficiency levels that are equivalent to the ELI's Level 3 and Level 4 respectively. Although the learners therefore comprised different cohorts who potentially had different instructors and RSA topics, these additional sources of variation are outweighed by the consistency that should result from only considering learners of comparable initial proficiency.

In order to determine initial proficiency and course placement, incoming ELI students are evaluated on three measures: the standardized Michigan Test of English Language Proficiency (MTELP, see Corrigan et al., 1979), and in-house listening and writing assessments. On the MTELP, possible scores range from 0 to 100. The 30 participants had an average score of 44.30, SD = 9.59, ranging from 25.00 to 74.00. On the in-house listening test, possible scores range from 0 to 25; here the mean score was 11.77, SD = 3.18, with scores ranging from 5.00 to 17.00. Finally, on the writing placement test, possible scores are between 0 and 5. The 30 participants' mean score was 2.50 with SD = .85, ranging from 1.00 to 4.00. The relatively high standard deviations illustrate a large degree of variation across learners' scores, but here I am concerned with whether learners in the two L1 groups have comparable initial proficiency. For this reason,

it is also worth considering initial proficiency by L1 by splitting the data. The ranges, means, and SDs for the two L1 groups are listed in Table 6.

Table 6. Initial proficiency scores by L1 for all 30 learners

L1	Placement Instrument	N	Minimum	Maximum	Mean	Std. Deviation
Arabic	MTELP	15	28.00	60.00	42.07	7.79
	listening test	15	5.00	17.00	11.20	3.57
	writing test	15	1.00	3.30	2.22	.71
Chinese	MTELP	15	25.00	74.00	46.53	10.91
	listening test	15	7.00	16.00	12.33	2.74
	writing test	15	1.00	4.00	2.78	.90

The means illustrate that on all three measures, the Arabic cohort had slightly lower placement test scores than their Chinese counterparts. Excluding learner C 611's data would give rise to the following table:

Table 7. Initial proficiency scores by L1 (C 611 excluded)

L1	Placement Instrument	N	Minimum	Maximum	Mean	Std. Deviation
Chinese	MTELP	14	25.00	74.00	46.79	11.27
	listening test	14	7.00	16.00	12.35	2.84
	writing test	14	1.00	4.00	2.72	.90

Even with learner C 611's data removed from the pool, the initial placement test scores exhibit the same trend.

An independent samples t-test failed to reveal a statistically reliable difference between the mean MTELP scores of Arabic learners and Chinese learners, with $t(28) = -1.291$, $p = .207$. In addition, an independent samples t-test was carried out on the listening test scores, and it also failed to find a reliable difference between the scores of Arabic and Chinese learners, with $t(28) = -.985$, $p = .338$. Finally, an independent samples t-test was performed on the writing sample scores and found no reliable difference between the Arabic and Chinese learners' scores, with

$t(28) = -1.888$, $p = .069$.[7] The fact that p approaches significance for the writing tests suggests that perhaps the Chinese learners were slightly more proficient writers upon beginning their ELI enrollment, but even if this difference were statistically significant, it may not have direct bearing on learner's oral performance, with which this thesis is concerned. Overall, the results of these t-tests allow us to infer that both groups of learners began their studies at the ELI with comparable initial proficiency and that any differences in performance and development must be attributed to other factors—such as L1.

4.3.2 Comparing grammatical functors in Arabic and Chinese

Luk and Shirai's (2009) review indicated that "learners can acquire a grammatical morpheme later or earlier than predicted by the natural order, depending on the presence or absence of the equivalent category in their L1" (p. 721). Therefore, it is worth considering potential correspondences of equivalent categories in the L1s of the learners under investigation in the current thesis. Recall that the nominal functors include plural *–s*, indefinite article *a/an*, and definite article *the*, while the verbal structures analyzed include regular past tense *–ed*, past irregular (*gave*, *went*), and third person singular present tense *–s*.

[7] These initial proficiency descriptive and inferential statistics were run with learner C 611's data included in the pool because he is included in the morpheme analysis. However, since he is excluded from the CAF analysis, independent samples t-tests were also run without his data. The scores still exhibit equality of variances and give similar results, failing to find a reliable difference between the scores of the two L1 groups (MTELP, $t(27) = -1.319$, $p = .198$; listening test, $t(27) = -.961$ $p = .345$; and writing test, $t(27) = -1.664$, $p = .108$. Thus, it is still fair to conclude that that the two L1 groups have comparable initial proficiency, regardless of whether C 611's data is included or excluded.

4.3.2.1 L1 Arabic

Arabic is a highly inflecting (fusional) Semitic language with VSO word order (Smith, 2001, p. 201) and is the first language of over 422 million speakers (Ribes, 2012). In addition to numerous dialects, there are two distinct main varieties of Arabic: Classical (a.k.a. Modern Standard) Arabic is the literary H variety used in writing and formal discussions and carries great prestige, while colloquial Arabic (a.k.a. Gulf Arabic) is the L variety used for everyday communication (Nydell, 2012, p. 94). Because all the subjects in this study are speakers of Gulf Arabic, the equivalent structures in this variety (hereafter Arabic) are considered, as it is likely their implicit knowledge of their L1 will influence English L2 output.[8]

Let us begin by comparing the equivalent nominal structures in Arabic. Arabic distinguishes number between singular, dual, and plural nouns, inflecting to mark the latter two. There are two types of plurals: sound plurals are formed by suffixing *–iin* to feminine nouns and *–aat* to masculine nouns with appropriate stem changes, while broken plurals are formed from the singular by changing the internal structure of the noun instead of adding suffixes. Because patterns are uncommon among broken plurals, most of the forms are irregular and must be learned on a case-by-case basis (Qafisheh, 1977, p. 105-6). For nouns following numerals greater than ten, Arabic uses a singular form, which could potentially transfer to English; for example, "I have ten brothers and sixteen *uncle" (Smith, 2001, p. 206). Based on Stockwell, Bowen and Martin's (1965) hierarchy of difficulty, this is a case of *coalescence* because the L1 Arabic dual and plural categories are realized only as plural in the L2. Stockwell et al. predict that coalescence is the easiest acquirable linguistic point after *correspondence*, in which there is a

[8] It is also plausible that Arabic learners' experience studying Classical Arabic, characterized by explicit metalinguistic knowledge, will affect their production of English, but a further investigation of this possibility is beyond the scope of this thesis.

one-to-one matching between a given form, category, or lexical item in the L1 and L2. Therefore, one can hypothesize that Arabic learners will be able to transfer their L1 representation of plurality to their English IL, but with some potential non-target forms.

Arabic has a definite article *al-* that attaches to nouns as a prefix but changes phonological form depending on environment. The definite article is used to refer back to previously mentioned indefinite nouns, for unique reference (i.e., *the sun*), and before a subset of proper nouns whose members are determined by lexical etymology (Qafisheh, 1977, p. 123). Smith (2001) notes that *al-* has a range of uses different from English: preceding all days of the week, some months in the Muslim calendar, and many names of towns, cities, and countries (p. 205-6). Arabic learners of English can therefore transfer their L1 knowledge of the definite article to English, but this might result in overgeneralizations (e.g., "he lives in *the India"; "we shop on *the Monday"). Arabic has no indefinite article, so English *a/an* is predicted to pose problems for L2 learners as this semantic category does not exist in their L1, and might result in developmental undersuppliance of *a/an*.

In terms of verbal structures, Arabic has a perfect tense that corresponds to English simple past (e.g., *he came*) and present perfect (e.g., *he has come*) (Qafisheh, 1977, p. 51). Depending on the verb stem, the perfect tense is formed either by adding inflectional suffixes that mark tense, person, number, and gender on sound verbs, or by changing the internal structure of the three- or four- consonant root (a.k.a. triliteral or quadriliteral) for weak verbs and doubled verbs. In general, the Arabic perfect tense signifies an action completed at the time of speaking, which corresponds to the semantics of the English past simple but will create difficulties when learners must *differentiate* between the past simple and present perfect, which comprise the same category in Arabic (Stockwell et al., 1965). However, for the past simple,

some degree of positive transfer of L1 knowledge can be expected, in which the suffixation of sound verbs can transfer to past regular *–ed*, and the internal changes of weak verbs can transfer to irregular past. Of course, this will be problematic given the mismatches between which verbs have regularly and irregularly formed past tense forms in Arabic vs. English. Finally, because Arabic verbs inflects for tense, person, number, and gender, learners should be able to transfer this representation to L2 English and mark regular third person singular verbs in the present tense with *–s*. However, some errors are expected because this is the only person/number combination in English that is marked by regular inflection in the present tense. In addition, this *–s* is always redundant because the obligatory subject already denotes person and number.

The equivalent categories and predictions regarding L1 influence are summarized in Table 8. Correspondence of form refers not to the phonological identity of a functor, but to whether it is a free form or a bound affix, and if bound, whether it is bound in the same place as in English.

Table 8. Arabic morphemes' potential for transfer

English functor	Correspondence of category	Correspondence of form	Expected transfer from L1 Arabic?
plural *-s*	+	+ for sound plurals - for broken plurals	+
definite article *the*	+	+	+
indefinite article *a/an*	-	-	-
regular past *–ed*	+	+ for sound verbs - for weak verbs	+
irregular past	+	+ for weak verbs - for sound verbs	+
third person singular present tense *-s*	+	+	+

+ indicates expected transfer; - indicates no expected transfer

4.3.2.2 L1 Chinese

Chinese is a Sino-Tibetan isolating SVO language that relies on word order (not inflection, as in Arabic) to mark grammatical relations and is structurally quite different from Indo-European languages like English. Chinese is a native language for one fifth of the world's population (Chang, 2001, p. 310) and is comprised of many dialects including Mandarin (a.k.a. Northern Chinese or *putonghua*, "common speech"), Cantonese, Wu, Xiang, Min, Hakka, and Gan (Po-Ching & Rimmington, 2004, p. xvii-xviii). Among the numerous varieties, I am concerned with Mandarin (hereafter Chinese), which is the basis for modern standard Chinese, the accepted written language for all Chinese (Chang, 2001, p. 310) and is the L1 of all of the participants.

According to Po-Ching and Rimmington (2004), with the exception of the restricted class of human nouns, under no circumstances do Chinese nouns inflect for case, gender, or number, and plurality must instead be inferred from discourse context or the category of the noun and its accompanying classifier (p. 1). In the case of human nouns, the suffix *–men* is sometimes viewed as a plural marker (Li, 1999); however, *-men* is highly restricted in its distribution. It normally relates to people in groups, and is used optionally in terms of address, e.g., *péngyoumen* 'friends'; *xiāngshengmen nǚshìmen* 'ladies and gentleman' (Lardiere, 2007, p. 199). It is also worth noting that the suffix *–men* can never be used with numbers (e.g., *sān ge háizi*, not **sān ge háizimen* 'three CL[9] children'). The most frequent and the only obligatory use of *-men* occurs in the closed class of personal pronouns (Lardiere, 2009, p. 194), listed in Table 9:

Table 9. Chinese personal pronouns

Person	Singular	Plural
1st	*wǒ*	*wǒmen*
2nd	*nǐ*	*nǐmen*
third	*tā*	*tāmen*

[9] CL = classifier

There has been great debate regarding the status of *–men* as a genuine, productive plural marker in Chinese (c.f. Li, 1999). Lardiere (2009) argues that *–men* is not a true plural marker but rather a collective marker based on five observations that: first, it is only obligatory on personal pronouns; second, its use on nouns other than humans is limited to pronouns; third, -*men* cannot be used with a quantifier or numeral, unlike languages with a true plural marker such as English and Arabic; fourth, when used with proper nouns, it may either have a plural or collective interpretation; and fifth, if a human noun is marked with *men*, there is an obligatory definite interpretation e.g., *háizimen* "the children," not "*(some) children" (Po-Ching & Rimmington, 2004, p. 10; see also Li, 1999, p. 81).

What is significant for this research is how L1 influence from Chinese will affect learners' production of English plural *–s*. The lack of L1-L2 correspondence is predicted to be problematic for learners, as quantified contexts such as *sān ge xuesheng* 'three CL students' prohibit use of *–men* in Chinese (**sān ge xueshengmen*) but are exactly the same contexts that require obligatory plural marking in English (Lardiere, 2007, p. 200). Likewise, Chinese nouns suffixed with *–men* are of an obligatorily definite reference, while English plural nouns are only definite if they co-occur with a definite determiner, e.g., *he knows three students* (indefinite) vs. *he knows the/those three students* (definite). In any case, even if *–men* is analyzed as a plural marker, as Li (1999) argues, it is more likely an element in Determiner than a regular plural in N such as English *–s* (p. 75). The lack of correspondence will require learners to reorganize and reassemble plural and definiteness features from their L1 to their L2 (Lardiere, 2007, 2009). Before complete feature reassembly occurs, one can expect developmental undersuppliance of plural marking in L2 English because number marking in the L1 is not obligatory beyond

personal pronouns (Lardiere, 2009, p. 198). In terms of phonology, Chinese does not allow complex codas, so learners may also omit plural *–s* for phonological reasons.

Unlike English, Chinese has no definite or indefinite articles (Chang, 2001, p. 318) and allows the free occurrence of bare arguments (Lardiere, 2009, p. 192). However, Chinese does have a deictic system of reference with demonstrative determiners, *zhè* "this" and *nà* "that," which are optionally followed by a classifier when used before a noun. But when a numeral precedes a noun, a classifier is obligatory. Huang, Li, and Li (2009) explain, "in the presence of numerals and demonstrative pronouns, a Chinese noun usually needs a classifier to specify the 'unit' with which the entities denoted by the noun are measured. Crucially, different nouns require different classifiers" (p. 14). Finally, although Chinese does not have an article system, the word *yī* "one" + CL + N is a construction comparable to English '*a/an* + N', but the fact that *yī* may be omitted attenuates the correspondence. In sum, the lack of an article system in Chinese should lead to undersuppliance of the definite article *the* and indefinite articles *a/an,* as well as oversuppliance of each in non-obligatory contexts (perhaps as the result of instruction; see Pica, 1983) resulting in overall low TLU.

In terms of verb forms, Chinese does not use inflectional tense markers. Instead, the concept of time is expressed by temporal adverbials, default viewpoint aspect, aspectual markers, modal verbs, or may be inferred from context (Lin, 2005, p. 2). Without an inflectional means of expressing tense or aspect in their L1, Chinese learners of English are expected to commit frequent errors of omission of regular past tense *–ed*; irregular past (e.g., *go, fly* instead of *went, flew*); and third person singular present tense *–s*. When the *–ed* and *–s* morphemes are realized at the end of a complex coda, as in *walked* and *talks,* these morphemes are additionally likely to be

omitted on a phonological basis because of the restricted syllable structure of Chinese, which only allows some nasals to occur in a coda (Juffs, 1990).

These correspondences (and lack thereof) as well as potential for transfer to L2 English are summarized in Table 9.

Table 10. Chinese morphemes' potential for transfer

English functor	Correspondence of category	Correspondence of form	Expected transfer from L1 Chinese?
plural *-s*	+/-	+	+/-
definite article *the*	-	-	-
indefinite article *a/an*	-	-	-
regular past *–ed*	-	-	-
irregular past	-	-	-
third person singular present tense *–s*	-	-	-

+ indicates expected transfer; - indicates no expected transfer

As compared to the L1 Arabic learners, much less L1 transfer from the Chinese group can be expected simply because Chinese is not an inflectional language.

4.3.3 Comparing learning styles and cultural influence

Let us now consider learning styles and communicative orientation based on cultural influence in terms of how they might affect the learners' oral production.

4.3.3.1 Arabic cultural influence

In Arabic, eloquence and rhetoric (including repetition and redundancy) are emphasized and admired far more than in Western cultures (Nydell, 2012, p. 97), and this sociolinguistic orientation is predicted to transfer to learners' acquisition of and performances in L2 English. Alabbad and Gitsaki (2011) note the disparity between learners' communicative goals and

outdated educational practices in many EFL classrooms in Saudi Arabia, which often emphasize grammar over oral communication skills. This mismatch is exemplified by the following quotation from an interview with a Saudi English language learner that illustrates the desired ability to successfully communicate:

> One of the main factors that has caused the current teaching method to be unsatisfactory is its over-focus on grammar and its neglect of the other language skills, like conversations… Giving the students chances for conversations and discussion is more important in learning than injecting the grammatical rules (Alabbad & Gitsaki, 2011, p. 15-16).

Other recent research confirms a disposition towards speaking in English. Juffs and Friedline's recent paper "Sociocultural Influences on the Use of a Web-Based Tool for Learning English Vocabulary" (2014) reports ESL students' answers to the question "What is the best way to learn new words?" Among the five Arabic-speaking students who answered this question, four suggested "oral interaction." Juffs and Friedline note the significance of such responses as compared to Korean-speaking students, most of whom

> mentioned reading or writing as a good way to learn vocabulary, but only three out of eleven mentioned speaking/oral skills. Overall, the Korean students mentioned study techniques that involved text (reading, writing, using dictionaries) rather than oral skills. In contrast, several Arabic-speaking students singled out reading and memorization as the worst way to learn new words. (p. 53)

Based on the cultural capital of eloquence and the inclination toward oral practice in the EFL/ESL classroom, Arabic speakers are expected to exhibit a high willingness to communicate in the L2 and to emphasize fluency over accuracy and meaning over form (Skehan, 1998).

4.3.3.2 Chinese cultural influence

Chinese learners' learning style and communicative orientation diverges greatly from that of Arabic learners and is more similar to the Korean learners interviewed by Juffs and Friedline

(2014). According to Chen, Warden, and Chang (2005), a crucial feature of the Chinese education system and culture in general is the importance of high school, university, and civil exams, for which memorization is the best way to prepare. Rote memorization is the favored learning strategy for Chinese characters, and is happily embraced for learning English as well. In fact, Chen et al. note that in China, "Books of English idioms are always big sellers, and many well-known and successful figures promote the ever popular memorize-a-dictionary strategy" (p. 625). In fact, Chang (2001) comments that many Chinese EFL and ESL students may spend considerable time on memorization at the cost of other kinds of interactional practice (p. 322).

Another factor to consider is the emphasis of written grammatical accuracy on high stakes standardized English examinations. Because there are no speaking assessments on such exams, oral proficiency and fluency in particular carry little cultural capital and utility for Chinese EFL learners. Furthermore, they have few to no opportunities for oral practice because there are so few native English speakers with whom to interact (Chen et al., 2005, p. 625), compounded by the scarcity of genuine English input in this EFL environment (p. 622). Therefore, unlike the fluency-oriented L1 Arabic learners, the Chinese subjects are more likely to prioritize accuracy and even complexity over fluency, with potential trade-off effects.

4.4 DATA

4.4.1 Data collection

In order to measure spontaneous oral output, this study considers Recorded Speaking Activities (RSAs), a formative assessment tool designed by and employed by the ELI in speaking class.[10] There are usually four RSAs per semester, where the first serves to introduce students to the assessment type, and the latter three activities are graded; this research only considers the graded RSAs. In general, the week before an RSA, the speaking instructor will present students with several possible topics. However, the precise prompt to which learners must respond is unknown until the RSA is performed.

RSAs are carried out in the computer lab following a set procedure. First, students are presented with a short prompt of one to four sentences and have one minute to plan their speech without taking notes or using any reference materials. Vercellotti (2012) notes that the planning conditions are most similar to the "no pre-task planning" groups described in the literature and, as such, reflect "pressured online planning" (p. 69). An example prompt for the topic *can't do here* is "Describe something that you liked to do when you were in your country but that you can't do here. Where did you do this? Why did you like it? How did it make you feel?"

Students then record themselves speaking for two minutes using a microphone. Next, they listen to their speech on headphones and transcribe exactly what they said, including fillers, self-corrections, and errors, thus offering them an opportunity to "notice the gap" between what

[10] All RSA data come from the Pittsburgh Science of Learning Center (PSLC) funded ELI Online Database.

they intended to say and what they actually said. Students then transcribe their errors and identify them as grammatical, vocabulary, or pronunciation based, allowing the learners to put their explicit knowledge of the L2 to use. They then report on such errors by recording a series of sentences of the form "I said *he go*. I should have said *he went*."

Learners are then given the opportunity to re-record their speeches, maintaining their original content but correcting as many of their errors as they can. The final RSAs are graded using an analytical rubric that provides separate scores for grammar, vocabulary, accuracy, and fluency and allows the instructor to provide each learner with not only a grade but also individualized feedback. No one element of the CAF triad is emphasized over the others in the task instructions or on the rubric.

McCormick and Vercellotti (2009) found that ELI students tend to focus on grammatical accuracy over fluency, lexical variety, or grammatical complexity when completing the self-correction aspects of the RSA, suggesting that their communicative orientation may focus on accuracy at the cost of complexity (syntactic and lexical) and fluency. Some learners are better than others at identifying their own errors; however, this self-awareness construct is beyond the scope of this thesis. Consequently, I only used the initial recorded speeches, considering students' transcriptions only when the intended utterance was unintelligible.

4.4.1.1 RSA topics and prompts

The RSA topics are general enough to allow all learners an equal opportunity to respond to a prompt completely and elaborately. Although the prompts vary by semester, many prompts are recycled over consecutive semesters, sometimes with minor adaptations (e.g., *pets* and *important event in my country* prompts). At least one prompt per semester is designed to elicit usage of the simple past tense. Appendix B includes a table of individual learners' RSA topics (B.1), as well

as the exact prompt that accompanied each topic (B.2). Table 11 contains a summary of the RSA topics and number of respondents at each level.

Table 11. RSA topics and number of respondents at each level

Topic	Number of respondents at Level 3	Number of respondents at Level 4
best friend	7	0
my country	3	0
funny or scary experience	10	4
favorite holiday	11	4
my background	3	0
important event in my country – A	4	0
a place you like	1	0
sports	3	0
upcoming vacation	3	0
next vacation	2	0
my city	3	1
first school	3	1
most important things	3	1
shopping for food	1	4
can't do here	1	4
custom in your country	1	4
country change past 50 years	2	0
complaint	0	1
favorite place	0	2
Pets – A	0	2
Pets – B	4	10
important person in my past	4	10
biggest problem in my country	4	10
free time	3	3
significant event	3	3
cultural differences	3	3
important event in my country – B	1*	3
important person in my country	1*	3
famous place	1*	3
learning English	0	1*
foreign language	0	1*
greatest accomplishment	0	2
local customs	0	1
problem	0	2
life pre-ELI	1	0
first day in Pittsburgh	1	0

a trip	0	1
university in my country	0	1
strategies to improve English	0	1
childhood	1	0
travelling in my country	1	0
confusing situation	1	0
job	0	1
vacation spot	0	1
renting	0	1

*Learner C 611's topics are indicated by an asterisk, denoting they are not included in the CAF analysis. The full prompt that accompanies each topic is listed in Appendix B.2.

Table 9 illustrates the heterogeneity of the topics to which students had to respond. Among the topics listed, the most frequent were those that concerned pets (N = 16); a funny or scary experience (N = 14); my favorite holiday (N = 15); an important person in my past (N = 13); and the biggest problem in my country (N = 14). It goes without saying that different prompts elicit different structures and vocabulary. For example, the *pets* prompts were either "How do people in your country feel about pets?" (N = 2) or "How do you feel about pets? Do many people have pets in your country? How are they treated in general?" (N = 14), both of which target simple present usage. In contrast, the prompts that ask students to "Talk about an important event that happened in the past in your country" or "Talk about a funny or scary experience that you had" elicit simple past usage. Since most prompts elicit personal or familiar information, no prompts should present an inherent advantage over any others for learners.

4.4.2 Transcription and coding

Oral production data is notoriously difficult to analyze, threatening the reliability and validity of some L2 studies of both performance and development (R. Ellis & Barkhuizen, 2005). For this reason, following Vercellotti (2012), I document all measures in this thesis. The raw CAF scores for all learners are included in Appendix C.

4.4.2.1 CAF analysis

The first step was to transcribe all 179 RSAs using PRAAT software, which allowed me to record and measure pauses, determining which were long enough to play a role in AS-unit divisions. Then, for the CAF analysis, learners' speeches were coded into clauses and AS-units following Foster et al.'s (2000) guidelines, discussed in Section 2.1.

Following Vercellotti (2012), I made a single modification to the AS-unit criteria regarding omitted copulas, illustrated by "__" in the following examples:

(11) C 127: | When he had a dog :: the dog __ just six months old |

(12) A 29: | And my family __ also there my father and my mother |

Foster et al. did not specifically grant such cases AS-unit status. However, omitted copulas frequently characterize the speech of Arabic and Chinese ESL students, likely due to the lack of a one-to-one corresponding category in their L1s, and the fact that copulas are not necessary to communicate the meaning of such messages. Vercellotti (2012) notes that utterances without a copula "function as an AS unit in the speech and have more meaning and complexity than a minor utterance, even without the copula" (p. 72). Thus, even if a VP lacks a copula, I still coded it as an AS-unit, following Vercellotti. Once the speeches were divided into clauses and AS-units, the syntactic complexity measure was calculated by dividing the number of total clauses by the total number of AS-units per RSA.

In terms of accuracy, errors were considered any deviations from TL norms that a speaker of American English generally would not produce (Lennon, 1991). Errors were categorized according to three subtypes: lexical, syntactic, and morphological. Examples of subtypes of each error are included by category in Table 12.

Table 12. Error typology and examples

Error type	Error subtype	Learner ID	Example
Lexical	wrong content word	A 12	Eid Alfeter comes after a month of **fastening***
		A 25	Also my father **learned*** me a good manner
	wrong part of speech	A 25	he is very **mercy***
	wrong content word wrong part of speech	C 127	in Pittsburgh some fish {some fish not is not} is not fresh | {he is, they is,} they are **freezer*** but not living.
	wrong preposition	C 126	**on*** my opinion I think :: feed pets {is like is} is better thing
Syntactic	article *a* omitted	C 127	that is * important event
	article *a* oversupplied	C 537	I am also busy :: when I have **a*** free time
	article *the* omitted	C 611	I think :: living in ***USA** is difficult because of ***language**
	article *the* oversupplied	C 611	**The*** Thailand's people are very kind
	omitted copula	C 520	and I * interested in computer
	oversupplied auxiliary	C 611	I'**m*** very like :: to eat this food
	omitted functional morphology	A 11	He likes * hangout
		A 12	He has been **advise*** me :: since I was a child
	omitted relative pronouns	C 633	And we choose the one :: * is available and flexible for us
	word order	C 914	Why mother's day **is*** my favorite?
	wrong tense	A 29	and when I came to American :: and I **see*** the dog :: I scared from that
	S-V number agreement	A 30	So we **was*** shocked :: when we heard that
	resumptive pronoun	A 163	when the one who I had an accident with **him***
Morphological	plural *–s* omitted	C 126	She has many best **friend***
	plural *–s* oversupplied	A 30	I was participating in everything every **activities***
	third sing. *–s* omitted	C 631	So everyone **need*** capable of using computer
	third sing. *–s* oversupplied	C 127	In my country China many people **likes*** badminton in the morning
	ill-formed past	C 126	For example {I pick choose} I **choosed*** my

	irregular oversupplied past *–ed*		first job
	omitted past *-ed*	A 11	I went to him :: to ask him about anything :: that I **want***
	possessive *–s* omitted	A 530	It is very crucial war in my country (1.2) | because this war was changed {people} Kuwaiti **people*** minds in many aspects
	possessive *–s* oversupplied	C127	And the people in Beijing want :: to welcome the **other's*** country's people :: to visit Beijing

*The error is marked by an asterisk and **bold** font when possible.

As discussed in Chapter 3, accuracy was calculated by dividing the number of error-free clauses by the total number of clauses in a given RSA.

Finally, fluency was measured in words per minute, including filled and unfilled pauses in the denominator. Words that comprised false starts, self-corrections, reformulations, and hesitation were not included in the numerator unless they comprised their own AS-unit, which was rare. Thus, this "global" measure of fluency reflects not only fluency speed, but also aspects of fluency breakdown and repair, as the latter two will also bring down the WPM measure.

4.4.2.2 Grammatical functor analysis

Past morpheme studies have considered accurate suppliance on up to 14 grammatical functors (Brown, 1973), but this was not possible given the small sample of data in the current study. Although the data were initially coded for 10 functors (plural *–s*; copula; auxiliary; progressive *–ing*; article *the*; articles *a/an*; past regular *–ed*; past irregular; third person singular *–s;* possessive 's), there were not enough instances of copula, auxiliary, progressive *–ing*, or possessive *'s* to perform a meaningful, systematic analysis. Therefore, I limited the study to the following morphemes: plural *–s*; articles *a/an* and *the*; past regular *–ed*; irregular past; and third person sing. present *–s*. However, because Krashen (1977) argued that the natural order only holds in studies with ten or more obligatory contexts per morpheme, I collapsed the data from the three

RSAs performed at Level 3 and the three RSAs performed at Level 4. Therefore, each learner has a Level 3 and a Level 4 accuracy score for each of the six morphemes, yielding 12 morpheme accuracy scores per learner.

As explored in Section 3, past research has relied on both SOC and TLU to measure accuracy on a given grammatical functor. SOC is beneficial because it assigns half credit to misformed but supplied functors, while an advantage of TLU is that it considers overgeneralization of morphemes in non-obligatory contexts, a hallmark of acquiring a given form in instructional settings (Pica, 1983). In order to get "the best of both worlds," I calculated morpheme accuracy scores as follows:

$$\text{(13) Specific accuracy score} = \frac{1 * (n \text{ correct suppliance in obligatory context}) + .5 * (n \text{ misformations in obligatory context})}{1 * (n \text{ obligatory contexts}) + 1 * (n \text{ suppliance in nonobligatory contexts})}$$

The misformations element is significant because here I included not only ill-formed functors (e.g., C 127 ***a** important holiday* for ***an** important holiday*) but also self-corrections. For example, Arabic learner 159 said, "{We check} we checked our mobiles" and although the initial formulation of "we check" is not included in the fluency word count so that it can reflect a drop in repair fluency, the production of past regular *–ed* on *checked* is only assigned half credit because the necessity of a self-correction should give rise to a lower accuracy score than if the morpheme were well-formed and did not need to be self-corrected in the first place.

In addition, any utterances that repeat all or part of the topic prompt are excluded from the morphemic analysis as learners may have been "primed" to supply certain morphemes if they just read them in context. For example, with the prompt "What is the biggest problem your country is facing today? How would you change it?", any repetition of "the biggest problem" was not coded for definite article *the.* Similarly, in the prompt "How do you feel about pets? Do

many people have pets in your country? How are they treated, in general?" any correct usage of *pets* was not considered in the plural count. However, if learners omitted a grammatical functor or oversupplied one when repeating the prompt, these omissions and oversuppliances were included in the specific accuracy score calculations.

Let us now review all the variables in this study before exploring the research questions and hypotheses.

4.4.3 Independent variables

The independent variables in this study include the learners' language background (Arabic or Chinese), gender, age at data collection, which semesters they were enrolled, RSA topics, years spent studying English in their homeland, years spent in an English environment, and their initial proficiency. Although years spent studying in their homeland and years in an English environment could potentially play a role in CAF and specific accuracy scores, I did not consider them in the statistical analysis because many learners' responses were unreliable. For example, just as Vercellotti (2012, p. 58, fn. 2) found, some learners reported five years of English learning in their homeland and three to five years of living in an English environment when answering a demographic survey at Level 3, but at Level 4, the same learners responded one to two years to both questions. Because initial proficiency was found to be comparable across the two groups of learners, I limit my statistical analysis to consider only L1 as an independent variable.

4.4.4 Dependent variables

The dependent variables are best divided into two subgroups: the CAF measures and specific accuracy on six grammatical functors.

In order to track global development, I operationalize CAF as explained in Chapter 2 and summarized in Table 13:

Table 13. Dependent variables through operationalization of CAF

Measure	Subtype	Operationalization
Syntactic complexity	Subordination	Average number of clauses / AS-unit
Global accuracy	Clausal accuracy	Average number of error-free clauses / total clauses
Global fluency	Speed	Words per minute • counting filled and unfilled pauses • not counting false starts, repetitions, reformulations

For simplicity's sake, I refer to these three dependent variables simply as complexity, accuracy, and fluency. Each learner will have scores for these three dependent variables at six observation times, yielding 18 total scores per learner. These raw scores are included in Appendix C.

As for the morpheme analysis, each learner has a specific accuracy score for each of the six grammatical functors including plural *–s*; articles *a/an* and *the*; past regular *–ed*; irregular past; and third person sing. present *–s*. The wide variety of topics elicited unequal uses of the six morphemes under investigation. It was thus necessary to collapse data from the three RSAs performed at Level 3, and the three RSAs from Level 4, following Spinner (2011) and as discussed in 4.4.2.2. This manipulation of the data yields six accuracy scores at two levels, giving a total of 12 scores per learner. The raw scores are included in Appendix D.

4.5 DATA ANALYSIS

4.5.1 Research questions and hypotheses

Two research questions and accompanying hypotheses were formulated to guide the investigation of L1 influence in the current research:

RQ1. Do L1 and cultural background influence the development of CAF over time?

H1. L1 and cultural background are predicted to have a significant interaction with CAF development over time, where Arabic learners have significantly higher initial fluency scores, while Chinese learners have higher complexity and accuracy scores, given the learners' different cultural backgrounds and communicative orientations. However, the Chinese learners are expected to make greater gains in fluency over time than their Arabic counterparts because the ESL environment is likely to invoke a shift in their distribution of attentional resources and an orientation that increasingly favors fluency.

RQ2. Is learners' accuracy on six grammatical forms influenced by their L1?

H2. Language background is predicted to have a significant effect on production of six grammatical functors based on whether or not corresponding categories exist in the learners' L1s. More specifically, Arabic learners are expected to have higher suppliance than Chinese learners on the following functors: plural *–s,* definite article *–the,* regular past *–ed,* irregular past, and third person singular *–s,* as these categories exist in Arabic but not Chinese. Both groups of learners are predicted to have similar undersuppliance of indefinite article *a/an,* as this category is absent from both Arabic and Chinese.

4.5.2 Statistical procedures

Mixed repeated measures (RM) ANOVA analyses are employed to answer both questions. Question 1 is addressed by a mixed between-within 2 (L1) x 3 (CAF) x 6 (time) RM ANOVA test, where CAF scores are transformed into Z-scores to ensure comparability of the three measures (following Larsen-Freeman, 2006).

Question 2 is investigated using a mixed between-within 2 (L1) x 6 (grammatical functor) x 2 (time) RM ANOVA test. In addition, two implicational scales (one at Level 3, one at Level 4) are used to answer Research Question 2 and explore the degree to which emergence of given grammatical functors fits into an implicational hierarchy, in which the acquisition of one functor implies the acquisition of one or more other functors for each learner (R. Ellis, 2008, p. 69). Coefficients of reproducibility and scalability are also calculated. Although I predict that language will influence morpheme accuracy, I do not hypothesize that learners' accuracy scores will adhere to a natural order, nor that Arabic learners will always place higher on the implicational scale than the Chinese learners (Spinner, 2011) despite L1 influence. I also hypothesize that the high degree of individual variability will lower the overall reproducibility and scalability of these implicational scales.

5.0 RESULTS

The results of this thesis are organized as follows. In Section 5.1, the results of the CAF measures are presented. Sections 5.1.1 – 5.1.3 contain descriptive statistics for each construct, while Section 5.1.4 contains inferential statistics for both L1 groups. The findings are discussed briefly in Section 5.1.5. In the following Section 5.2.1, I present the descriptive statistics of the grammatical functor analysis, dividing the functors into nominal and verbal categories. The functors are then ranked in Section 5.2.3. In 5.2.4, two implicational scales are presented, followed by inferential statistics in Section 5.2.5. The results are briefly discussed following the presentation.

5.1 CAF MEASURES OF GLOBAL DEVELOPMENT

5.1.1 Complexity development

Complexity was operationalized through syntactic complexity by subordination, measured in the average number of clauses per AS-unit. The mean scores and standard deviations for all learners and by L1 per observation point are presented in Table 14.

Table 14. Complexity by L1 at six observation points

Observation Point	L1	Mean	Std. Deviation	N
1	Arabic	1.403	.36429	15
	Chinese	1.351	.14892	14
	Total	1.378	.27814	29
2	Arabic	1.483	.29790	15
	Chinese	1.577	.27315	14
	Total	1.529	.28516	29
3	Arabic	1.499	.24407	15
	Chinese	1.664	.32169	14
	Total	1.579	.29122	29
4	Arabic	1.707	.43361	15
	Chinese	1.616	.41468	14
	Total	1.663	.41953	29
5	Arabic	1.846	.37670	15
	Chinese	1.885	.39276	14
	Total	1.865	.37813	29
6	Arabic	1.953	.54332	15
	Chinese	1.916	.43786	14
	Total	1.935	.48678	29

A visual inspection of these descriptive data reveals that for all 29 participants, there was a tendency to improve over time, with the average complexity beginning at 1.378 clauses per AS-unit and increasing to 1.935 by observation point 6. However, there is a high degree of variation, with the SD as large as .49 at point 6. At observation points 2, 3, and 5, the Chinese learners exhibited higher average complexity, while the Arabic learners did at points 1, 4, and 6. These means are represented visually in Figure 2.

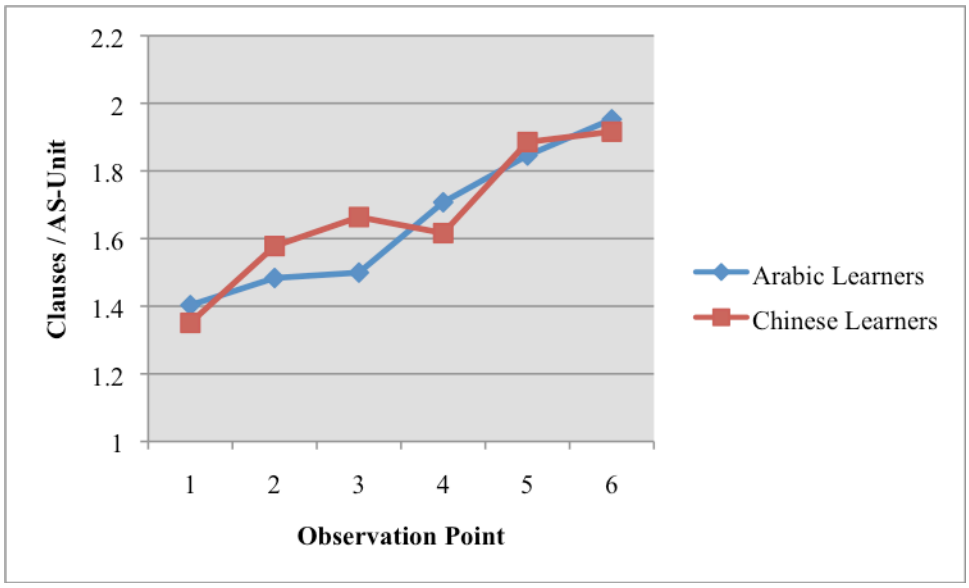

Figure 2. Mean complexity by observation point and L1

5.1.2 Accuracy development

Accuracy was measured in the number of error-free clauses divided by total clauses. Table 15 illustrates that the Arabic group had slightly higher mean accuracy than the Chinese learners at all observation points except point 6, but that the differences at each observation point were relatively small given the high degree of variation, with standard deviations up to .21.

Table 15. Accuracy by L1 at six observation points

Observation Point	L1	Mean	Std. Deviation	N
1	Arabic	.549	.14627	15
	Chinese	.511	.13620	14
	Total	.531	.14032	29
2	Arabic	.518	.16058	15
	Chinese	.494	.12987	14
	Total	.507	.14445	29
3	Arabic	.487	.15512	15
	Chinese	.436	.16232	14
	Total	.462	.15792	29
4	Arabic	.584	.16692	15
	Chinese	.464	.21404	14

	Total	.526	.19734	29
5	Arabic	.591	.12788	15
	Chinese	.538	.16853	14
	Total	.566	.14864	29
6	Arabic	.560	.15843	15
	Chinese	.561	.13877	14
	Total	.560	.14660	29

Unlike syntactic complexity by subordination, there was not a marked tendency to improve in accuracy over time. This change over time is represented visually in Figure 3. Both groups of learners' global accuracy scores decline from observation point 1 to 3 and increase from point 3 to 5. From point 5 to 6, the average Arabic scores decline, while they grow for the Chinese learners. Therefore, the changes seen in accuracy scores are far from linear.

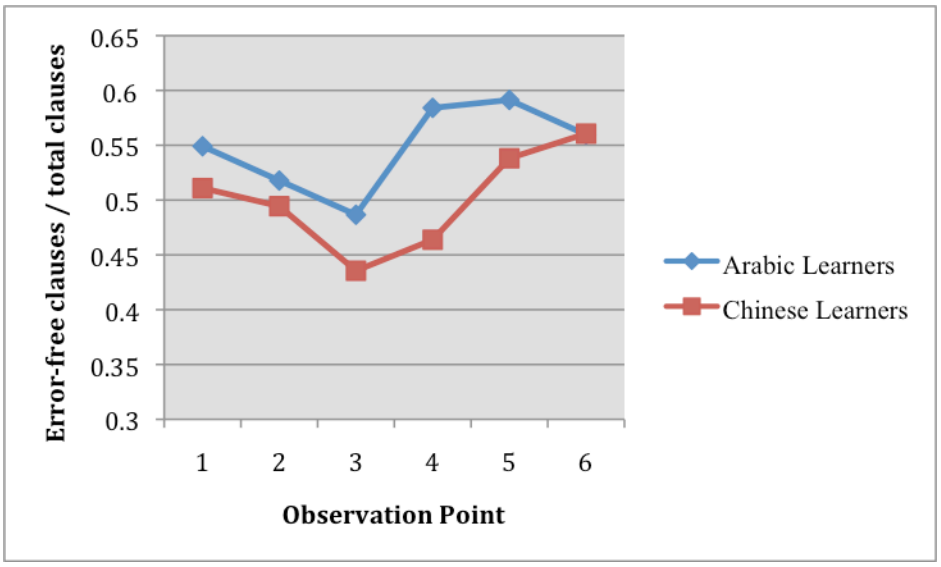

Figure 3. Mean accuracy by observation point and L1

5.1.3 Fluency development

Fluency was measured in words per minute, including all filled and unfilled pauses in the denominator, and subtracting all false starts, non-rhetorical repetitions, hesitations, and prior

formulations of self-corrections from the denominator. Table 16 illustrates the mean fluency scores over time. There was not linear growth over time in the group total.

Table 16. Fluency by L1 at six observation points

Observation Point	L1	Mean	Std. Deviation	N
1	Arabic	71.30	28.60908	15
	Chinese	50.89	12.78972	14
	Total	61.45	24.35122	29
2	Arabic	70.43	23.82902	15
	Chinese	54.79	18.96692	14
	Total	62.88	22.67564	29
3	Arabic	72.35	21.68805	15
	Chinese	57.14	17.22823	14
	Total	65.00	20.80467	29
4	Arabic	70.87	17.52986	15
	Chinese	59.36	19.21540	14
	Total	65.31	18.95638	29
5	Arabic	78.47	22.77597	15
	Chinese	64.66	17.64077	14
	Total	71.80	21.28810	29
6	Arabic	75.44	22.19625	15
	Chinese	63.22	11.48210	14
	Total	69.54	18.60517	29

If the two L1 groups are compared, it is clear that the Arabic learners had higher fluency rates than the Chinese learners at all observation points, despite high standard deviation values. The larger SDs for Arabic learners are likely due to learner A 30, whose initial fluency of 138.3 WPM is over double that of some other learners. However, the group trend remains, which can be attributed to to cultural influence and the value placed on oral proficiency in Arabic culture. The difference between the two groups of learners' fluency scores in especially evident in Figure 4.

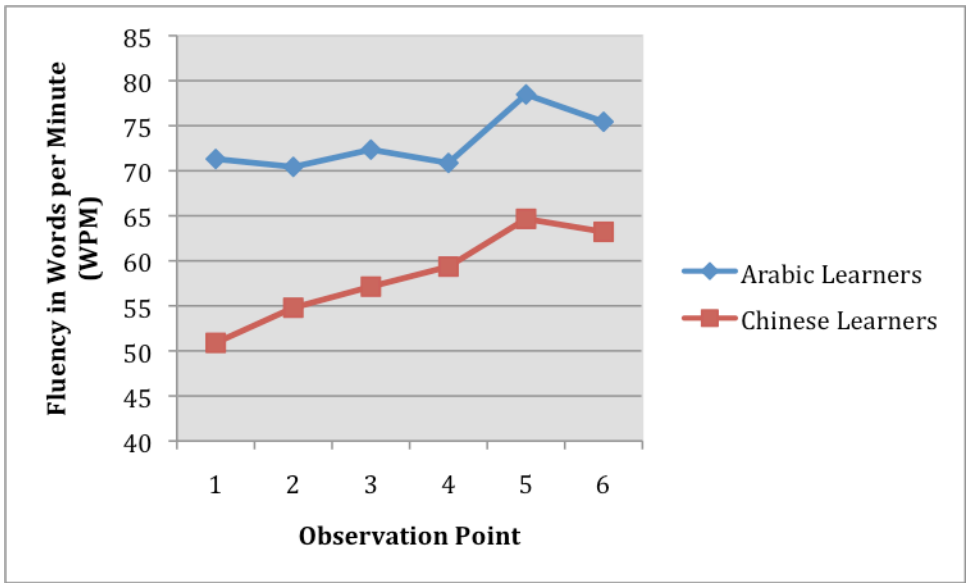

Figure 4. Fluency in WPM by observation point and L1

In terms of change, Figure 4 illustrates that the Chinese learners' fluency tended to increase over time, with the only decrease occurring between observation points 5 and 6. Although the Arabic learners' scores also decreased between points 5 and 6, they did not exhibit a tendency to improve linearly over time and their fluency only grew from point 4 to 5. This growth was preceded by a relatively steady mean fluency from point 1 to 4.

Because each CAF measure occupies a different scale, the performance measures were transformed to z-scores to ensure comparability across the different indices, following Larsen-Freeman (2006). The estimated marginal means for the CAF measures for the two L1 groups are illustrated in figures 5 and 6.

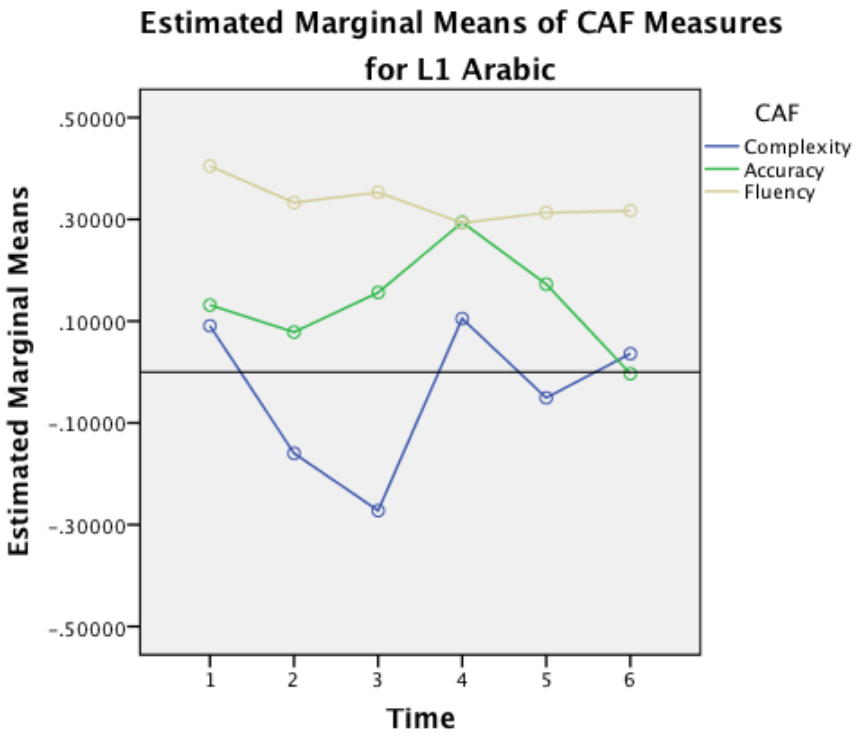

Figure 5. CAF measures for L1 Arabic learners over time

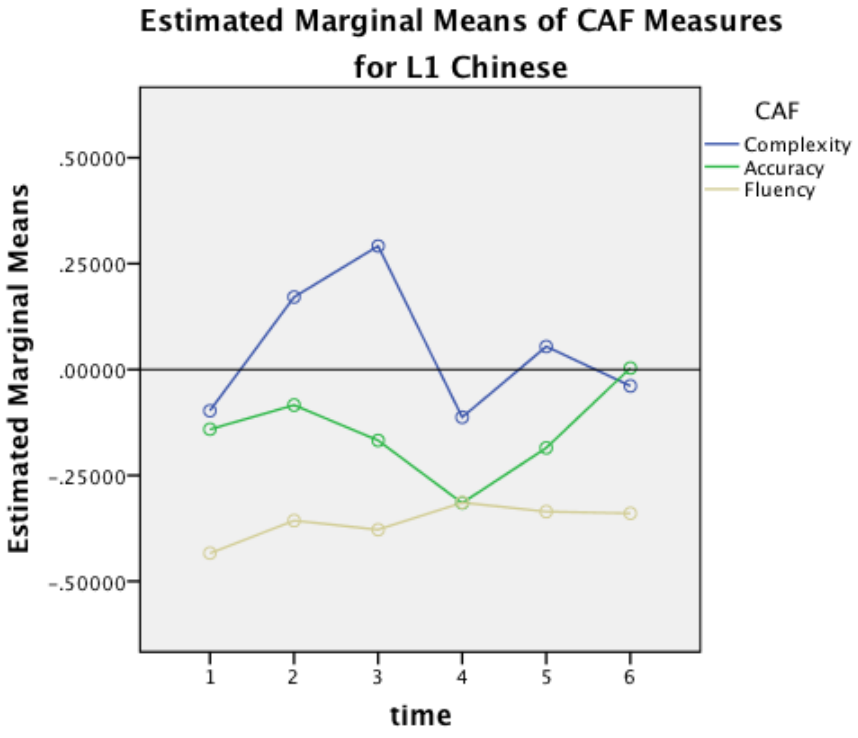

Figure 6. CAF measures for L1 Chinese learners over time

A comparison of Figures 5 and 6 indicates that at all six observation points, the Arabic learners had much higher fluency rates and slightly higher accuracy scores (not at point 6) than the Chinese group. However, grammatical complexity oscillated, with neither group exhibiting a clear advantage over the other. These charts, when viewed from an individual differences perspective, suggest that the different L1 groups of learners have different priorities and thus allocate their limited attentional resources differently. The Arabic learners seem to prioritize fluency over accuracy and accuracy over complexity, while the Chinese learners exhibit the opposite hierarchy, with higher complexity than accuracy scores, and still lower fluency scores when compared to the overall group means.

5.1.4 Inferential statistics

First, a mixed between-within 2(L1) x 3 (CAF) x 6(time) repeated measures Analysis of Variance (RM ANOVA) was performed. It revealed no overall between-groups effect for L1, with $F(1,27) = 2.56$, $p = .121$, partial eta squared = .087, observed power = .338. This is likely because of the large degree of intra-individual variation, illustrated by high standard deviation values and confirmed by a glimpse at Appendix C.

To investigate within-subjects effects, I first ran Mauchly's test of sphericity, which tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix (SPSS). Although the results of Mauchly's test for sphericity were not statistically reliable, I followed Larson-Hall's (2010) advice to take Howell's (2002) suggestion of using the Greenhouse-Geisser correction to degrees of freedom for analyzing statistical effects. Therefore, all reported inferential statistics were derived using the conservative Greenhouse-Geisser correction.

The RM ANOVA test revealed that there is a significant interaction between the measure CAF and the L1, with $F(1.946, 52.537) = 3.63$, $p = .034$, partial eta squared = .119, observed power = .638. The effect size was not larger because of the high degree of variation. Yet this finding confirms that L1 interacts reliably with CAF measures because the Arabic learners had considerably higher fluency and slightly higher accuracy than the Chinese learners.

The RM ANOVA showed no main effect for time with $F(4.367, 117.913) = 0$, $p = 1.0$, partial eta squared = .00, observed power = .05, again because there is so much intra-individual variation across performances. Similarly, it follows that there was no significant effect for the interaction of time with L1, with $F(4.367, 117.913) = .48$, $p = .79$, partial eta squared = .017 observed power = .166. The same is true for the interaction between time and CAF, which also failed to exhibit a statistically significant effect, with $F(6.862, 185.283) = 0$, $p = 1.0$, partial eta squared = .00, observed power = .05. Finally, there was no significant effect for the interaction of time with CAF measure with L1, with $F(6.862, 185.283) = .63$, $p = .73$, partial eta squared = .023, observed power = .262. Importantly, the lack of effects for interactions between time and CAF and/or L1 illustrates the lack of linear growth among individuals. Although the group averages may indicate improvement over time, this trend may not apply to any individual learner.

5.1.4.1 Arabic learners' CAF development

Given the significant effect for an interaction between CAF measure and L1, separate RM ANOVA tests were run for each L1 group to look more closely at change in CAF. The Arabic learners showed no significant effect for time, with $F(3.963, 55.479) = 1.49$, $p = .22$, partial eta squared = .096, observed power = .429. This is likely due to the high degree of intra-individual variation even among just the Arabic learners.

Next, I ran separate RM ANOVAs for the three CAF measures, and the results tell a different story. For the complexity measure, there was a main effect for time, $F(3.069, 42.961) = 5.34$, $p = .003$, partial eta squared = .276, observed power = .914, indicating a tendency to improve over time. For the accuracy measure, there was no main effect for time, with $F(3.069, 42.965) = 1.26$, $p = .302$, partial eta squared = .082, observed power = .315. Similarly, for the fluency measures, there was no main effect for time, $F(3.963, 55.488) = 1.33$, $p = .27$, partial eta squared = .087, observed power = .387. These results indicate that for the Arabic learners, the only significant improvement over time occurred in complexity, and not in fluency or accuracy.

5.1.4.2 Chinese learners' CAF development

In contrast, the RM ANOVA on the Chinese learners' data showed a significant main effect for time, with $F(3.118, 40.532) = 3.41$, $p = .025$, partial eta squared = .208, observed power = .835.

Again, the CAF measures were transformed into z-scores to ensure comparability across them. Separate RM ANOVAs were run for the three CAF measures and the results revealed just where development was significant. For the complexity measure, time had a significant effect, with $F(3.425, 44.519) = 5.84$, $p = .001$, partial eta squared = .310, observed power = .954. For the accuracy measure, there was no main effect for time, with $F(3.990, 51.876) = 1.82$, $p = .14$, partial eta squared = .320, observed power = .629. Finally, for the fluency measure, time had a significant effect, with $F(3.103, 40.336) = 3.16$, $p = .033$, partial eta squared = .196, observed power = .700.

5.1.5 Discussion of results

In this discussion, I would like to touch briefly on three main points: the high degree of individual variability, development over time by L1, and some methodological implications.

The mixed RM ANOVA revealed that there was a significant effect for the interaction between L1 and CAF measure, but no main effects for L1 or for time. This finding is attributed to the large high degree of individual variability, which creates great variance in the data, indicated by high standard deviation scores. Specifically, individual learners failed to improve linearly over time, often regressing in CAF measures from one data collection point to the next. Such high variability is consistent with Larsen-Freeman's (2006) characterization of L2 as a CAS on the verge of a phase shift. However, it is worth noting that the observed group effects may not correspond to any one individual's path of development, and thus the data must be considered in closer detail.

When I analyzed the L1 groups separately, I found that there was no main effect for time for the Arabic learners on any of the three CAF measures. After I ran separate analyses for each measure, I found a significant effect for time only on complexity, which tended to improve over time. In contrast, there was no main effect for time on the global accuracy or fluency measures. On the contrary, when the Chinese learners' CAF measures were measured, a significant main effect was found for time. Breaking this down by measure, there were significant effects of time in the complexity scores (like the Arabic learners) and in the fluency scores (unlike the Arabic learners). Finally, like the Arabic learners, there was no main effect for global accuracy over time.

The increases in complexity by subordination across both groups of learners are likely due to the effects of being in an IEP and the grammatical growth that occurs in an instructed

SLA environment over a relatively long period of time. However, the analysis of the fluency scores indicated a main effect for only the Chinese learners, and not the Arabic participants. I attribute this difference to their divergent cultural orientations. Because oral proficiency in English is somewhat neglected in Chinese culture but emphasized in Arabic culture, the Chinese learners had more room for improvement, while the Arabic learners may have been closer to their ceilings upon initial enrollment.

Finally, both groups of learners' accuracy scores failed to exhibit a significant interaction with time. However, this finding is not surprising (cf. the lack of accuracy gain in study abroad programs), given the crudeness of this global accuracy score. There is no difference because all the errors are combined together, with no distinction made between an error on a simple structure or on a more complex construction. Furthermore, all errors are weighted equally, be they morphosyntactic and attributable to L1, universals, or idiosyncratic to individual learners. For this reason, in order to better track development, it is necessary to analyze errors on a smaller scale because the grain size of analysis may make a difference in charting individual development. This observation leads to the next analysis: that of specific accuracy on six grammatical functors.

5.2 GRAMMATICAL FUNCTOR ACCURACY

In terms of specific accuracy scores, it is best to first consider accuracy by dividing the six functors under investigation into nominal and verbal categories. The nominal functors include the plural *–s* morpheme, definite article *the*, and indefinite article *a/an*. The verbal functors are the regular past *–ed* morpheme, irregular past verb forms (*went, left,* etc.), and the third person

singular present tense *–s* morpheme. Raw accuracy scores by learner by level are reported in Appendix D.

5.2.1 Nominal functors' specific accuracy scores

The nominal functors' mean specific accuracy scores and standard deviations for all learners and for learners by L1 are presented in Table 17.

Table 17. Nominal functors' mean specific accuracy scores by L1

Level	Grammatical Functor	L1	Mean	Standard Deviation	N
Level 3	plural *–s*	Arabic	.7167	.187773	15
		Chinese	.4544	.154317	15
		Total	.5856	.215196	30
	the	Arabic	.7069	.187347	15
		Chinese	.5476	.218871	15
		Total	.6273	.215940	30
	a/an	Arabic	.5819	.197245	15
		Chinese	.5693	.244603	15
		Total	.5756	.218420	30
Level 4	plural *–s*	Arabic	.7348	.158720	15
		Chinese	.4807	.239657	15
		Total	.6077	.237866	30
	the	Arabic	.7193	.134797	15
		Chinese	.6605	.159998	15
		Total	.6899	.148400	30
	a/an	Arabic	.6174	.189801	15
		Chinese	.5796	.247541	15
		Total	.5985	.217582	30

When both groups of learners are combined, the following hierarchy of accuracy at both Level 3 and 4 emerges: *the* (62.73% at Level 3; 68.99% at Level 4) > plural *–s* (58.56% at Level 3; 60.77% at Level 4) > *a/an* (57.56% at Level 3; 59.85% at Level 4). Furthermore, a comparison of the scores across levels indicates that overall, learners improved more on *the* than accurate usage of plural *–s* or *a/an*.

When learners are divided by L1, at both levels, the Arabic learners had higher specific accuracy scores than the Chinese learners on all three nominal functors; however the difference is larger for plural *–s* than either article. Moreover, the high standard deviation scores indicate a large degree of variability in accuracy scores. Let us look briefly at how this variability manifests itself by looking at boxplots for the three nominal functors by level.

Figure 7 contains boxplots of specific accuracy scores for the three nominal functors at Level 3, dividing scores by L1 group.

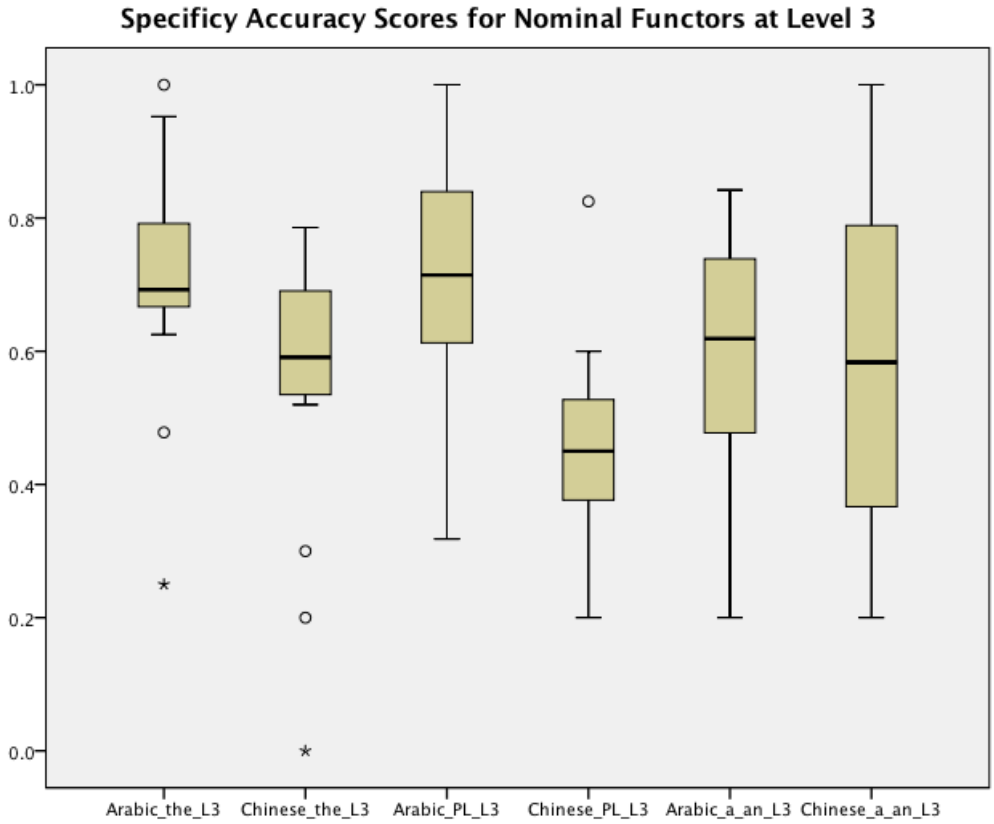

Figure 7. Mean accuracy scores for nominal functors by L1 at Level 3

The leftmost two columns illustrate that for the definite article *the*, Arabic learners have 29.08% higher mean accuracy (70.69%) than the Chinese learners (54.76%), which is consistent with hypotheses based on a corresponding definite article category in Arabic but not in Chinese.

However, there is a large degree of variability at Level 3, evident in the handful of outliers in Figure 7. For *the* accuracy, outliers include Arabic learners A 159, A 404, and A 481, and Chinese learners C 631, C 126, and C 157. The relatively large number of outliers in a small sample size demonstrates the high degree of variation in scores.

The two middle columns illustrate specific plural accuracy by L1 at Level 3. Here one can see a larger range of scores among the Arabic learners, but overall, they are 57.72% more accurate (71.67%) than the Chinese learners (45.44%). In fact, Chinese outlier C 298's high plural accuracy score is closer to the Arabic mean and would suggest that this participant is acting more like an Arabic learner. In any case, the large difference between the L1 groups' plural suppliance is consistent with hypothesis 2 that Arabic learners would be more accurate, as a comparable plural category exists in their L1 but not in Chinese.

Finally, for the indefinite article *a/an*, the rightmost two columns demonstrate that the Arabic learners had 2.22% higher mean accuracy scores (58.19%) than the Chinese learners (56.93%), but with a larger degree of in-group variation among the Chinese. The fact that neither group outperforms the other is consistent with my hypothesis, based on the fact that the indefinite article category exists in neither L1.

At Level 4, see similar trends are evident, shown in Figure 8.

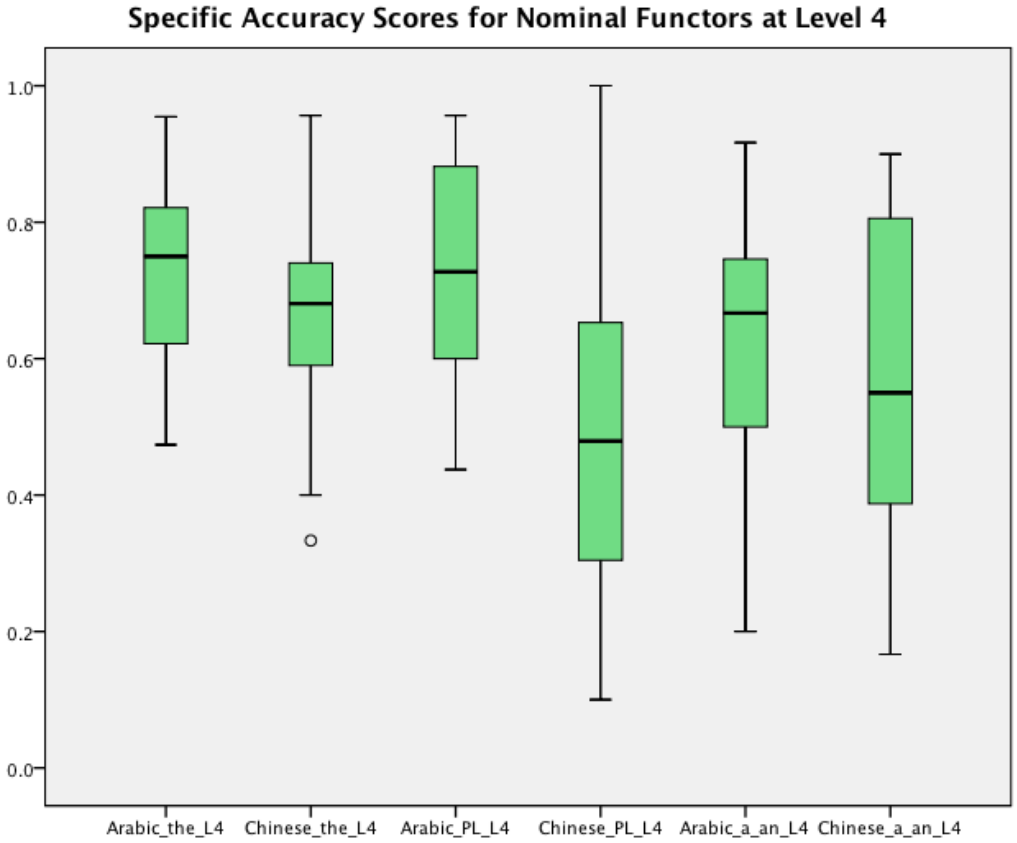

Figure 8. Mean accuracy scores for nominal functors by L1 at Level 4

In terms of articles, again the Arabic learners are slightly more accurate than the Chinese learners, but with both groups exhibiting considerable variation in the range of scores. For the definite article *the*, the two left-most columns illustrate that the Arabic group was 8.893% more accurate (71.93%) than the Chinese learners (66.05%); but recall that at Level 3, the difference was much larger, at 29.08%. I will address this point in the discussion section. This time, Chinese learner C 611 is an outlier with his low suppliance of *the*. For the indefinite articles *a/an*, the rightmost two columns of the chart show that the Arabic learners were 6.51% more accurate (61.74%) than the Chinese learners (57.96%). This is slightly higher than the 2.22% difference between the groups' means at Level 3, but the large range of *a/an* accuracy scores at both levels obviate an analysis of these differences before considering inferential statistics.

Finally, for plural *-s*, again the Arabic learners were 52.85% more accurate (73.47%) than

the Chinese (48.07%). This margin of difference between the group means is similar to that at Level 3, where the Arabic learners were 57.72% more accurate. Such a result is consistent with the hypothesized large L1 effect that would be evident here. The fact that the difference between group means gets slightly smaller (as also seen with *the*) would suggest that L1 influence plays a larger role at lower levels of proficiency. The attenuation of L1 effects could also be attributed to the instructed SLA environment in which learners find themselves.

5.2.2 Verbal functors' specific accuracy scores

Let us now turn to the verbal functors under consideration in this study. The mean accuracy scores and standard deviations for regular past *–ed*, irregular past, and third person singular present *–s* at Level 3 and Level 4 are presented in Table 18.

Table 18. Verbal functors' mean specific accuracy scores by L1

Level	Grammatical Functor	L1	Mean	Standard Deviation	N
Level 3	regular past *–ed*	Arabic	.5605	.354682	15
		Chinese	.4378	.350763	15
		Total	.4991	.352165	30
	irregular past	Arabic	.4565	.309130	15
		Chinese	.5144	.264921	15
		Total	.4855	.284401	30
	third person sing. present *-s*	Arabic	.0806	.141269	15
		Chinese	.1222	.284986	15
		Total	.1014	.222017	30
Level 4	regular past *–ed*	Arabic	.5070	.257831	15
		Chinese	.5519	.388463	15
		Total	.5295	.324750	30
	irregular past	Arabic	.6072	.237894	15
		Chinese	.6082	.363324	15
		Total	.6077	.301741	30
	third person sing. present *-s*	Arabic	.4917	.501041	15
		Chinese	.2033	.308491	15
		Total	.3475	.434322	30

When both groups of learners are combined, the following hierarchy of accuracy emerges at Level 3: regular past *–ed* (49.91%) > irregular past (48.55%) > third singular *–s* (10.14%). At Level 4, the hierarchy changes to irregular past (60.77%) > regular past *–ed* (52.95%) > third singular *–s* (34.75%). It is immediately clear that on average, learners had the largest gains on third singular *–s*, followed by irregular past, and only improved slightly on the regular past *–ed*.

If one compares accuracy scores by L1 group, at Level 3, the Arabic learners had higher scores than the Chinese learners on regular past *–ed,* while the Chinese learners were slightly more accurate with irregular past and third person singular present *–s*. At Level 4, the trend changes, with the Arabic learners considerably more accurate on third person singular present *–s*, while the Chinese learners were more accurate on regular past *–ed*; their scores for irregular past were comparable. However, the differences between the two groups means on many of these measures are small, and the standard deviations are large, demanding a visual inspection of the variation. Therefore, let us now look at a handful of boxplots.

Figure 9 contains boxplots of specific accuracy scores for the three verbal functors at Level 3, dividing scores by L1 group.

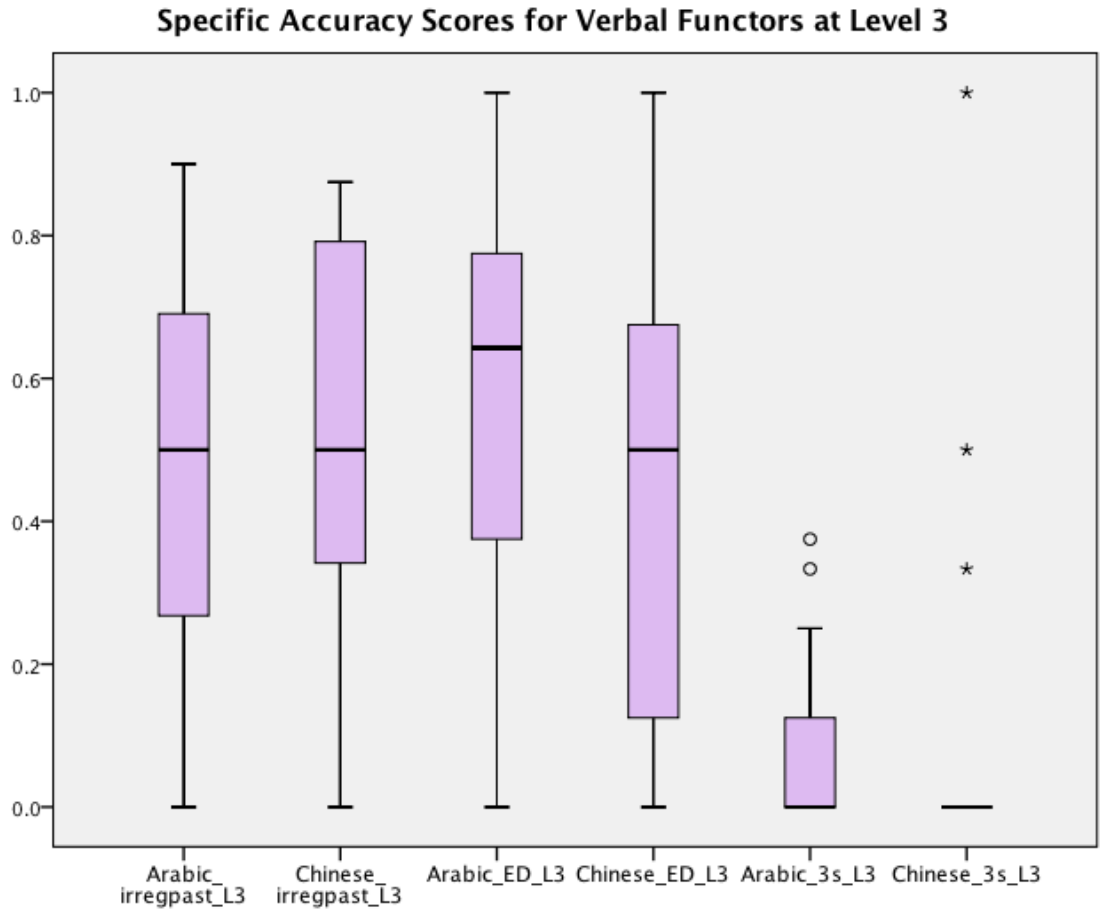

Figure 9. Mean accuracy scores for verbal functors by L1 at Level 3

The leftmost two columns illustrate accuracy on irregular past verb forms. Despite similar median scores, on average, the Chinese learners were 12.70% more accurate (51.44%) than the Arabic learners (45.65%), but both groups of learners' scores ranged considerably, and the high degree of variation precludes reading too deeply into this between-group difference. Next, on regular past *–ed*, the middle two columns indicate that the Arabic learners were 28.028% more accurate (56.05%) than the Chinese learners (43.78%), a much larger difference than the one associated with irregular past forms. But again, there was a considerable degree of variation in responses indicated by the large range of accuracy scores (from 0% to 100%) for both L1 groups. Based on L1 influence, I had predicted that the Arabic learners would be more

accurate on both regular and irregular past forms, as an equivalent to the simple past is formed via inflection and stem changes in Arabic but not in Chinese; however, this prediction was only borne out by the Level 3 data for the regular past.

For third person singular present tense *–s*, the right-most two columns show that both groups had median scores of 0, and the group averages were very similar, with the Arabic learners 8.06% accurate on this form, and the Chinese learners, 12.22% accurate. Before trying to explain this difference, however, note that Arabic learners A 65 and A 12 and Chinese learners C 301, C 282, and C 177 were all outliers, with higher accuracy on this form than the rest of the group. In fact, for the Chinese learners, all but the three aforementioned outliers had a specific accuracy of 0% on third singular *–s*; therefore, the differences indicated by the group means are misleading. Instead, usage of this specific verbal inflection seems to be highly influenced by individual variation.

Before drawing any conclusions or attempting to explain these between-group differences, let us look at the accuracy on the three verbal functors at Level 4, shown in Figure 10.

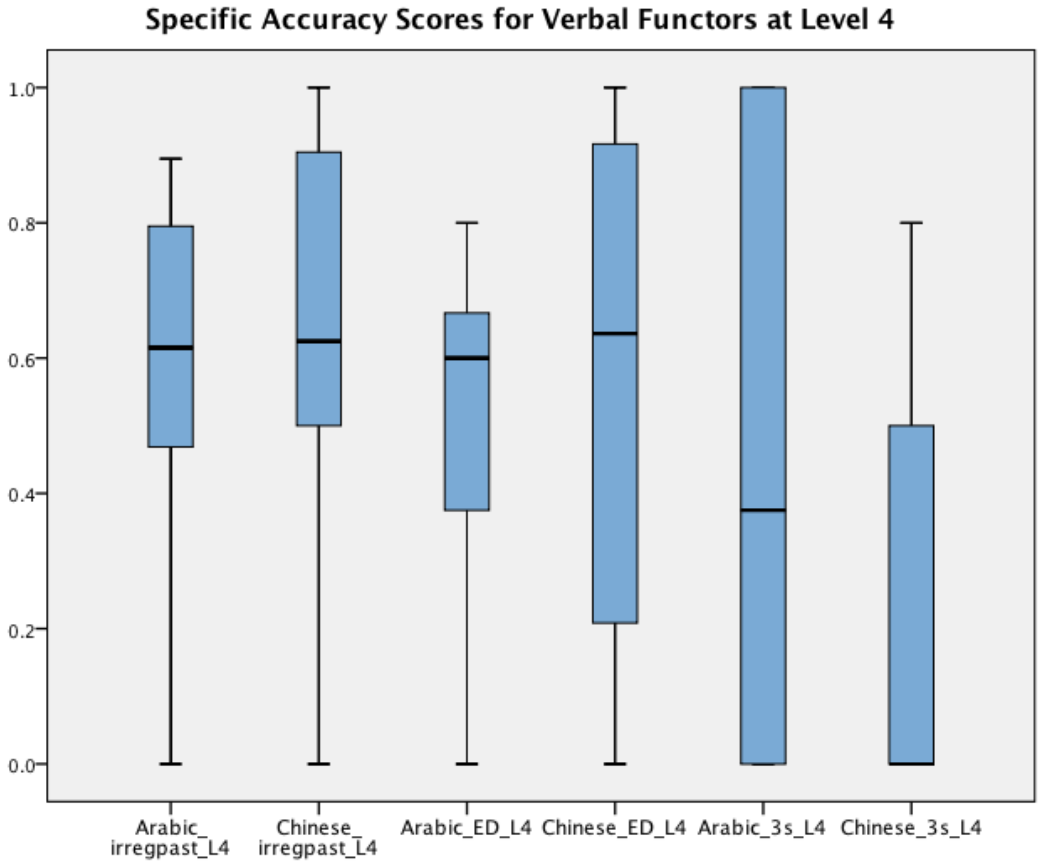

Figure 10. Mean accuracy scores for verbal functors by L1 at Level 4

Beginning with the irregular past in the leftmost two columns, the two L1 groups have strikingly similar means and medians, with Arabic learners on average 60.72% accurate, and Chinese learners 60.82% accurate. However, the boxplot shows that the Chinese learners had a slightly higher range of scores. As compared to the Level 3 scores, on average the Arabic learners had a slightly greater gain in accuracy on irregular past, increasing by 15.07%, while the Chinese learners' mean score only grew by 9.37%. Yet there was more variation (illustrated by the 50% percentile occupying a larger range) for both groups at Level 3 than Level 4.

Next, the two middle columns indicate that for regular past *–ed*, again the Chinese learners were 8.85% more accurate (55.19%) than the Arabic learners (50.70%). Yet the Chinese learners' accuracy scores exhibited greater variability, while the 50% percentile of Arabic

learners' scores occupied a smaller range. Note that this is the opposite trend of that exhibited in Level 3, where the Arabic learners were 28.03% more accurate than the Chinese. This result of a higher Chinese mean on regular past *–ed* at Level 4 goes again the predictions outlined earlier. From Level 3 to 4, the Chinese learners' mean accuracy increases by 25.07%, while the Arabic learners' mean accuracy decreases by 9.54%, an unexpected divergence.

Finally, for third person singular present *–s*, the two rightmost columns show two striking results. The first is the large range of responses among both groups of learners, but especially the Arabic students, whose 50% percentile occupies the whole range of possible accuracy scores. Second, the Arabic learners have a mean accuracy score (49.17%) that is 141.82% larger than the Chinese mean accuracy score (20.33%). This between-groups difference is much larger (and occurring in the opposite direction) than the one observed at Level 3.

5.2.3 Ranking of grammatical functors

Past studies have suggested both a natural order of acquisition and of accuracy for grammatical morphemes (Krashen, 1977) regardless of L1. If the mean nominal and verbal scores for all learners are combined, the following accuracy hierarchies at Level 3 and 4 emerge.

(14) Level 3: *the* > plural *–s* > *a/an* > reg. past *–ed* > irreg. past > third sing. *-s*

(15) Level 4: *the* > plural *–s* > irreg. past > *a/an* > reg. past *–ed* > third sing. *-s*

From Level 3 to Level 4, irregular past moves two spots up on the hierarchy, as both groups of learners—but especially the Arabic students—made significant progress on these lexically stored forms. If there were a natural order, one would expect the hierarchy to be the same at both levels of enrollment, with individual accuracy scores increasing but the ranking of grammatical

functors remaining the same. This is clearly not the case, but it is unclear whether this is due to L1, instructional effects, and/or other sources of individual variation.

If L1 does play a significant role in accuracy of grammatical functors, then one could expect to see different orders for the Arabic and Chinese learners, but the same hierarchies within each group at both levels. If the role of being in an instructed SLA environment were the sole determining factor, then one could expect to see the same changes in hierarchies across Level 3 and 4 for both groups of learners. The actual hierarchies are as follows:

(16) Arabic Level 3: plural *–s* > *the* > *a/an* > reg. past *–ed* > irreg. past > third sing. *-s*

(17) Arabic Level 4: plural *–s* > *the* > *a/an* > irreg. past > reg. past *–ed* > third sing. *-s*

(18) Chinese Level 3: *a/an* > *the* > irreg. past > plural *–s* > reg. past *–ed* > third sing. *-s*

(19) Chinese Level 4: *the* > irreg. past > *a/an* > reg. past *–ed* > plural *–s* > third sing.*-s*

Overall, there are very different hierarchies for the two L1 groups, most evident in the higher ranking for plural *–s* for Arabic learners at both levels. This suggests that L1 is playing a role, but the effect must be confirmed by inferential statistics (see Section 5.2.5).

In terms of within-group changes by L1 group, the orders shift by level for both groups of learners. For Arabic learners, irregular past moves up a ranking from Level 3 to 4, suggesting that the effect of exposure in an English environment is responsible for greater gains in irregular past accuracy than any of the other grammatical functors, as irregulars are more frequent than regular past forms. For the Chinese learners, even more changes in ranking by level are visible. For example, at Level 3 the Chinese learners were, on average, more accurate on *a/an* than *the*, but this trend is reversed in Level 4. In addition, irregular past and regular past *–ed* both move up one position from Level 3 to Level 4, suggesting that enrollment in an IEP has particularly evident effects for accurate production of the simple past tense. Finally, plural *–s* accuracy

moves down one position in the ranking. Thus, although the Chinese group improves in average plural accuracy from 45.44% at Level 3 to 48.07% at Level 4, this change is smaller than the gains made in *the*, irregular past, and regular past *–ed*. Clearly, the accuracy exhibited by all 30 learners on the six morphemes is conditioned by a combination of factors that include, but are not limited to, L1 influence, other individual differences, and specific development while in an instructed SLA environment (operationalized through time).

5.2.4 Implicational scales

Implicational scales or scales allow us to find structure in variability and often demonstrate that what seems like free or random variation is actually significantly constrained. More specifically, sociolinguist John R. Rickford (2002) explains "In linguistics … implicational scales depict hierarchical co-occurrence patterns in the acquisition or use of linguistic variables by individuals such that *x* implies *y* but not the reverse. When linguistic variables are distributed in implicational patterns, the scope of variability is significantly constrained" (p. 143).

In this case, I am investigating the degree to which grammatical functors exhibit an implicational hierarchy in which acquisition (or accurate usage) of one functor implies accuracy on others. For example, if learners can accurately produce third singular *–s*, then they can also use plural *–s* accurately; however, the reverse is not true. Implicational scales are especially useful because each individual's accuracy on each morpheme at both levels is considered separately. To make better sense of the descriptive statistics and explore whether these grammatical functors exhibit an implicational hierarchy, I created two implicational scales—one at Level 3, and one at Level 4—following the procedures described by Spinner (2007, 2011).

5.2.4.1 Level 3 implicational scale

The Level 3 implicational scale contains columns for each of the six grammatical functors, arranged in descending order by the total mean accuracy scores. Each row represents an individual participant.

In order to make sense of the accuracy scores and put the participants in an order, it was necessary to set a threshold that determines the productivity of emerging morphological markers and syntactic elements. Following Young-Scholten, Ijuin, and Vainikka (2005), I determined productivity by setting a 60% threshold in which a form must be used accurately in at least 60% of all obligatory contexts to be considered "emerged." Then, following Spinner (2011), I considered each learner's scores on all six morphemes. If an accuracy score was greater than or equal to 60% with three or more tokens, then the learner's score was marked with ** to denote emergence. However, very often there were a limited number of tokens and contexts at Level 3 or 4. If this was the case, then, as Spinner (2011) writes, "that result should be considered tentative" (p. 542) and such cases of ≥ 60% on only one or two tokens are denoted with one asterisk, *. Next, if learners had less than 60% accuracy on a grammatical functor and there were three or more tokens, the score is marked with N. If accuracy is less than 60% with only one to two tokens, then the score is marked with -N-, but again, such a result is tentative. Finally, if there were no instances of the structure (either supplied or omitted) at a given level, the score is marked with /.

To determine a ranking of individuals, I tallied the number of asterisks (** or * are each worth one point) per learner and sorted the learners in ascending order by total number of asterisks, indicating emergence of a functor. Following Hatch and Farhady (1982), learners with the same tally of asterisks may be rearranged so as to reduce the number of errors in the Table. In

addition, any case of / allows the learner to potentially be moved to the next group of learners in order to reduce error. For example, a learner with three asterisks and one empty cell (/) may be moved into the group of learners with four asterisks (Spinner, 2001, p. 544). The implicational scale for all 30 learners at Level 3 is presented in Table 19.

Table 19. Level 3 implicational scale

Learner ID	*the*	*a/an*	plural *–s*	past *–ed*	irregular past	third sing. *-s*
C 988	N	N	N	-N-	N	/
C 611	N	N	N	N	N	-N-
C 631	N	N	N	-N-	N	-N-
C 537	N	N	-N-	**	N	/
C 456	**	N	N	-N-	N	/
C 270	**	N	N	/	N	N
C 633	**	N	N	N	N	-N-
A 163	**	**	N	N	N	N
A 25	**	**	N	-N-	N	N
C 914	N	N	N	**	**	N
C 177	N	N	**	N	**	N
C 282	N	N	**	N	**	-N-
A 481	N	N	**	N	N	-N-
A 12	**	**	**	N	N	N
A 129	**	**	**	N	N	/
A 159	**	**	N	N	**	-N-
A 30	**	**	N	**	N	-N-
C 298	**	**	**	-N-	-N-	N
A 530	**	**	N	**	-N-	/
A 65	**	**	N	*	-N-	N
C 520	**	N	**	-N-	**	-N-
A 29	**	N	N	**	**	/
A 404	N	**	**	*	/	N
C 126	N	N	**	**	**	N
C 127	**	**	**	**	-N-	N
C 301	**	N	**	*	N	*
A 11	**	N	**	**	**	N
A 45	**	**	**	/	**	N
A 157	**	**	**	**	**	-N-
A 241	**	**	**	**	**	/

** = emerged with ≥ 60% accuracy with 3 or more tokens
* = emerged with ≥ 60% accuracy with 1-2 tokens
N = not emerged with < 60% accuracy with 3 or more tokens
-N- = not emerged with < 60% accuracy with 1-2 tokens
/ = no instances of the structure in the corpus at Level 3

Previous researchers have relied on two simple calculations to determine whether the set of data corresponds to a true order of emergence: the Coefficient of Reproducibility (C of R) and the Coefficient of Scalability (C of S) (Spinner, 2007, 2011).

The C of R is calculated based on the number of "errors'" or deviations in the table and is used to establish how predictable the results are for all learners at Level 3. It is calculated as follows (Guttman, 1944; Hatch & Farady, 1982; Spinner 2011):

(20) C of R = 1 – (total number of errors / total number of cells)

A C of R that exceeds .9 generally implies a predictable pattern (Hatch & Farady, 1982; Rickford, 2002). However, since the data come from spontaneous oral production, it is also possible that some learners have neither tokens nor contexts of a given functor (indicated by /), resulting in empty cells. In order to maximize the power of scaling, one must consider that such empty cells weaken the validity of the results (Rickford, 2002, p. 157); therefore, I modified the denominator to count only filled cells. At Level 3, the total number of errors (Ns on the left and ** or *s on the right of the boundary line) was 36, and the total number of filled cells is 170 (30 individuals x 6 functors = 180 total – 10 empty cells). Therefore, at Level 3, the C of R = .788 [1 – (36/170)]. This value indicates that the grammatical functors do not exhibit a predictable pattern, and the results of individual students cannot be reliably predicted using this order of acquisition (Spinner, 2011, p. 547).

Next, the C of S is a figure "intended to indicate if a set of variables is truly scalable (i.e., a developmental pattern)" (Spinner, 2011, p. 544). Following Spinner (2007), the C of S is calculated by first finding the Minimal Marginal Reproducibility (MMR) by dividing the total number of asterisks by the total number of responses. Then, the Percentile Improvement in

Reproducibility (PIR) is calculated by subtracting the MMR from the C of R. Finally, the C of S is calculated as follows (Spinner, 2007, p. 141-2):

(21) C of S = PIR / (1-MMR)

Hatch and Farady (1982) write that the C of S should be over .6 or .65 if the table is to be considered truly scalable. In this case, the MMR = .435 [74 asterisks /170 total responses]. The PIR = .353 [.788 - .435]. Finally, the C of S = .625 [.353 / (1 - .435)]. Since this figure just barely exceeds the .6 threshold, it indicates a scalable table. However, just because the grammatical functors are scalable does not imply that the results are reliable. In fact, the low C of R shows that individual students' results cannot be reliably predicted using this order of acquisition.

On a final note, if L1 were the only conditioning factor, then one would expect to see learners with the same language background clustering together, with all the Chinese participants at the top of the table (because of lower accuracy due to lack of L1 correspondence), and all the Arabic learners at the bottom. Clearly, this is not the case, and an individual's position on the table cannot be determined by L1 alone. Let us now see whether a similar lack of clustering and unreliable C of R is also found at Level 4.

5.2.4.2 Level 4 implicational scale

The order of columns in the Level 4 implicational scale is different from that at Level 3, based on the divergent ranking of overall accuracy explored in Section 5.2.3, which I argue is due to the instructed SLA environment. The process employed for sorting individuals and determining ranking by counting the total number of asterisks is that same as that described in Section 5.2.4.1. The Level 4 implicational scale is presented in Table 20.

Table 20. Level 4 implicational scale

Learner ID	*the*	plural –*s*	irregular past	*a/an*	past -*ed*	third sing. -*s*
C 611	N	N	-N-	N	N	/
C 298	N	N	-N-	**	-N-	**
C 537	N	N	-N-	N	*	-N-
A 481	N	**	N	N	N	-N-
C 988	**	N	/	N	/	-N-
C 270	**	N	-N-	N	-N-	-N-
C 127	**	N	**	N	N	-N-
A 25	**	N	-N-	**	N	N
C 914	**	N	**	N	**	/
C 631	**	N	**	N	**	/
C 177	N	**	**	**	-N-	N
C 126	**	**	**	-N-	-N-	-N-
A 45	**	**	-N-	**	-N-	-N-
C 633	**	N	**	N	**	/
A 241	**	**	**	N	**	-N-
A 157	**	**	**	N	N	*
A 404	**	**	**	-N-	N	**
A 30	**	**	**	**	N	N
A 163	**	**	N	**	**	-N-
A 11	N	**	**	**	**	-N-
C 456	**	N	**	**	*	-N-
A 29	**	N	**	**	**	-N-
A 159	**	**	N	**	**	*
A 129	**	**	N	**	**	**
C 301	**	**	N	**	**	**
C 520	**	**	*	**	**	/
A 12	**	**	**	**	N	*
C 282	**	**	**	**	**	-N-
A 65	**	**	**	N	**	*
A 530	**	**	**	N	**	*

** = emerged with ≥ 60% accuracy with 3 or more tokens
* = emerged with ≥ 60% accuracy with 1-2 tokens
N = not emerged with < 60% accuracy with 3 or more tokens
-N- = not emerged with < 60% accuracy with 1-2 tokens
/= no instances of the structure in the corpus at level 4

At Level 4, there were seven empty cells, meaning 173 cells were filled. Among these 173 cells, the total number of errors was 30. Therefore the C of R = .827 [1 – (30/173)], which

again falls below the .9 threshold. This low C of R indicates that the grammatical functors do not exhibit a predictable pattern, and consequently, this order of acquisition cannot be used to reliably predict the results of individual students. Next, I calculated the C of S by following the steps outlined above. The MMR = .578 [100 asterisks / 173 filled cells]. The PIR = .249 [.827-.578]. Finally, the C of S = PIR / (1-MMR) = .590 [.249 / (1 - .578)]. This C of S falls below Hatch and Farady's (1982) threshold; therefore it indicates that the table is not truly scalable.

Again, the learners do not cluster together by language background any more at Level 4 than at Level 3, suggesting that L1 is not the only factor determining a learner's accurate production of these six grammatical morphemes.

The C of R and C of S for the implicational scales at both levels 3 and 4 are summarized in Table 21.

Table 21. C of R and C of S at Levels 3 and 4

Measure	Level 3	Level 4
Coefficient of Reproducibility (C of R)	.788	.827
Coefficient of Scalability (C of S)	.625	.590

Although the Level 3 coefficient of scalability is significant at the 60% threshold, it does not hold up to the stricter 65% threshold. Therefore, I treat it as insignificant and do not explore any implications of the potential scalability of the table at Level 3. More important is the low coefficient of reproducibility at both levels, indicating that the results are not predictable for all learners.

5.2.5 Inferential statistics

In order to look more systematically at the effects of L1 and enrollment in an IEP (i.e., time) on grammatical functor accuracy, I performed a mixed between-within 2 (L1) x 6 (grammatical functor) x 2 (time) repeated measures Analysis of Variance (RM ANOVA). An overall between-groups effect for L1 was found, with $F(1,28) = 4.45$, $p = .044$, partial eta squared = .137, observed power = .531, determining that there is a reliable effect for L1. Based on the descriptive statistics, this L1 difference is most evident in the scores for plural *–s* and *the* at Level 3, and plural *–s* and third person singular *-s* at Level 4, with Arabic learners having higher scores than the Chinese participants on these four measures.

To investigate within-subjects effects, Mauchly's test of sphericity was performed. The results of the test for sphericity were statistical; for functor, Mauchly's $W(14) = .150$, $p = .000$; and for functor by time, Mauchly's $W(14) = .367$, $p = .025$. Consequently, I again took Howell's (2002) suggestion to use the Greenhouse-Geisser correction to degrees of freedom for analyzing statistical effects. Therefore, all reported inferential statistics were derived using the conservative Greenhouse-Geisser correction.

The RM ANOVA test revealed that there is a significant main effect for time, with $F(1,28) = 15.01$, $p = .001$, partial eta squared = .349, power = .962. This result indicates that speakers do tend to improve over time, with a significant difference in their accuracy scores at Level 3 vs. Level 4. Next, the effect for the interaction between time and functor approaches significance, with $F(.579, 100.90) = 2.21$, $p = .080$, partial eta squared =.073, observed power = .599. Similarly, the main effect for the interaction between time, grammatical functor, and L1 also approaches significance, with $F(.555, 100.90) = 2.12$, $p = .091$, partial eta squared = .070, power = .578. Although the interactions between (a), time and functor, and (b), time, functor and

L1 are not statistical at the p = .05 level, the fact that they both approach significance suggests that these variables' interaction should not be neglected. In other words, the effects of being in an instructed SLA environment are close to having a statistically significant effect, and the observed power is larger here than it was for the CAF measures' interaction with time. These results are discussed briefly in Section 5.2.6.

5.2.6 Discussion of results

Unlike the global accuracy measure employed in the CAF analysis of the data, here L1 effects are much more visible in learners' accuracy scores because of the smaller grain size employed to explore the data.

First of all, on plural *–s*, the Arabic learners were significantly more accurate than the Chinese learners at both Level 3 and 4. The difference between their mean accuracy scores was relatively stable over time. I attribute this difference to the fact that plurality exists in Arabic and is marked either by an inflectional suffix on sound nouns or by changing the internal structure of the noun for broken plurals. In contrast, Chinese has no productive plural marker, and the closest equivalent, the suffix *–men* is more like a collective marker that is highly restricted in its distribution and interacts with definiteness in a different way than English plural *–s*. Therefore, the hypothesis was supported by the data.

In terms of articles, the Arabic learners were more accurate than the Chinese learners on definite *the* at both levels (as hypothesized), yet the difference between the Arabic and Chinese mean accuracy was much smaller than the difference between their plural scores. It is also worth noting that there was a considerably larger difference between their group means at Level 3 than at Level 4. This finding would suggest that L1 exerts a larger influence either at lower levels.

Overall, these findings are consistent with the hypothesis, but the effect is smaller than predicted based on the fact that there is a corresponding definite article category in Arabic but not in Chinese. Perhaps the reasons that Arabic learners do considerably better than the Chinese on plural *–s* but only slightly better on *the* is due to mismatches in appropriate usage of the definite article in Arabic vs. English. Definite article usage in English is subject to pragmatic constraints (Hawkins, 1991), while plural is not. For the indefinite article *a/an*, I predicted that both groups of learners would have similarly low accuracy scores (as compared to *the*) because the indefinite article category exists in neither L1. This is exactly what the data illustrate. The Arabic learners were only slightly more accurate, likely because they have a definite article category in their L1.

Next, let us turn to the verbal functors. In distinguishing between regular and irregular past forms, it is worth noting that they represent different types of knowledge and processing. The regular past *–ed* morpheme requires syntactic processing to be applied correctly, while the irregular past requires lexical processing since these are memorized forms. In the case of the learners here, the Arabic group was actually less accurate on regular past *–ed* at Level 4 than Level 3, suggesting that perhaps their syntactic representation was undergoing restructuring, a point discussed further in Section 6.2.2. In contrast, the Chinese learners reliably increased in accuracy on *–ed* from Level 3 to Level 4.

For the irregular past, the Chinese learners were slightly more accurate than the Arabic group at both levels, but the difference was more marked at Level 3 than Level 4. I attribute this to cultural L1 influence and the emphasis placed on memorization as a learning technique in Chinese culture. The only way to produce irregular past forms accurately is to memorize them and access lexical knowledge quickly and efficiently. Perhaps the learning style of Chinese students in their homeland makes them more able to accurately supply irregular past forms

during on-line production, while the Arabic learners are instead focused on fluency (as explained in the CAF results).

Finally, the third person singular present tense *–s* responses exhibited the most variation and least consistency of all six grammatical functors. At Level 3, only a handful of learners (7 total) produced this form at all. Their frequent omission cannot be due to purely phonological factors, because if this were the case, then one could expect to see similar accuracy scores for plural *–s* and third sing. *–s*, as both have the same allomorphy. But this is clearly not what the data suggest. Instead, I attribute the frequent omission of third person *–s* to its redundancy, which N. C. Ellis (2006) argues is a significant factor in SLA, and plays a much larger role in L2 than in L1 acquisition. The third person *–s* morpheme will never encode necessary referential information about the subject—because English is not a pro-drop language, subjects are always included, and thus this third person verbal morphology is entirely redundant to communicating referential information. In contrast, in the absence of a quantifier, determiner, or numeral, plural *–s* is often the only way to transit referential information about number.[11] Despite the redundancy of the third singular *–s* inflection, learners do tend to make progress from Level 3 to Level 4. At Level 3, only six of the 30 learners (three of each L1) supplied this inflection, while at Level 4, 13 individuals (eight Arabic, five Chinese) used this morpheme, with varying levels of specific accuracy. Thus, despite its redundancy, learners are able to develop accuracy on this form thanks to their enrollment in an instructed SLA environment. However, it is strange that the Arabic learners make greater gains than the Chinese on this functor if L1 influence is supposed to be attenuated as proficiency increases. I address this paradox in Chapter 6.

[11] See Young (1993) for an in-depth discussion of redundancy in interlanguage and its relation the functional hypothesis.

On a final note, the learners did not cluster together on the implicational scales by L1, indicating that L1 alone does not determine learners' specific accuracy on the six grammatical functors. This result is important because it reminds us that L1 influence, though statistically significant, cannot be overestimated as the *only* factor contributing to between-individual differences. For this reason, in Chapter 6, I discuss other sources of inter- and intra-individual variation and explore their significance with respect to the results found by this thesis. But before beginning the discussion, as a side note, I would like to mention two examples of L1 influence.

5.2.6.1 Possessive *'s* morpheme

Although my quantitative analysis did not systematically investigate what percentage of learner errors could be directly attributed to L1, the coding schema allows for a basic exploration of L1 effects. One such example is accurate usage of possessive *'s*, a construction that has a comparable structure in Chinese but not in Arabic. Luk and Shirai (2009) found that Japanese, Korean, and Chinese learners acquired possessive *'s* earlier than comparable Spanish learners of English and attributed this difference to correspondence (or lack thereof) of this category in learners' L1. Although this morpheme was originally included in my coding, there were not enough instances (either supplied tokens or omissions) to allow a meaningful statistical analysis. However, by taking the 79 RSAS performed by the 15 Chinese learners and the 80 RSAs from the 15 Arabic learners, some simple calculations can still be performed. For the Chinese group, I found 27 tokens of correct possessive *'s* suppliance (excluding repetitions of the prompts such as "Describe an important event in your *country's* history"), three omissions, and four cases of oversuppliance, giving rise to a specific accuracy score of 79.41%. For the Arabic learners, there were only four tokens of correct possessive suppliance, with one misformed usage and three omissions, giving rise to a specific accuracy score of 56.25%. Thus, not only are the Chinese

learners more accurate on this morpheme, but they use it *much* more readily than the Arabic learners, likely due to a parallel structure in their L1 that is absent from Chinese (Kamimoto, Shimura & Kellerman, 1992).

5.2.6.2 Relative clauses

Another place to look for qualitative L1 effects is in the construction of relative clauses, which are formed differently in English, Arabic, and Chinese and are notoriously difficult for English language learners (Gass & Selinker, 2008). Arabic relative clauses feature no relative pronouns, but have obligatory resumptive pronouns that reflect filler-gap relations, so Arabic learners' production of relative clauses are predicted to be marked by the omission of relative pronouns and inclusion of resumptive pronouns (Schachter, 1974) if L1 influence is at play. Consider learner A 11's utterance from his third Level 3 RSA: "My favorite holiday is Eid Alfeter. It comes after a month __ we fast in *it," where the omitted relative pronoun (which is optional in English since it is an object) is underscored and the incorrect resumptive pronoun is marked by an asterisk. In contrast, relative clauses in Chinese tend to function like adjectival modifiers, preceding the noun they modify with a relative marker *de* at the end (Po-Ching & Rimmington, 2004) but are much less frequent than relative clauses in English. Because of the lack of a comparable construction in their first language, Chinese participants are expected to use fewer relative clauses altogether (Kamimoto et al., 1992) but for their production to be marked by omission of the relative pronoun, such as learner C 177's first Level 4 RSA: "Actually, there are a lot of people in Taiwan __ have pets."

An analysis of the 80 RSAs from the Arabic learners reveals a total of 24 resumptive pronouns and 15 omitted relative pronouns, while one resumptive pronoun and 14 omitted relative pronouns are present in the Chinese learners' 79 RSAs. These figures indicate that L1

influence absolutely affects production of relative clauses in English, especially evident in the Arabic learners' more frequent erroneous inclusion of resumptive pronouns. Overall, these findings are consistent with N. C. Ellis' (2006) conclusion that "difficulties of adult L2 acquisition are a result of prior L1 learning, entrenchment and transfer" (p. 185).

6.0 SUMMARY AND GENERAL DISCUSSION

This discussion chapter is divided into three sections. In Section 6.1, I explore global development in CAF measures and the significance of the inferential statistical results. In Section 6.2, I turn to the specific accuracy measures and what they might mean. Section 6.3 contains a more general discussion of the implications of this thesis.

6.1 GLOBAL DEVELOPMENT IN CAF MEASURES

In order to track development of complexity, accuracy, and fluency and the role of L1 in ESL students' spontaneous oral production, I gathered and coded six consecutive RSAs performed over eight months as 30 learners progressed from a low intermediate to high intermediate level of proficiency. Complexity was operationalized as the average number of clauses per AS-unit per RSA; global accuracy was the number of error-free clauses over total clauses; and fluency was the average number of words per minute.

The mixed RM ANOVA I ran did not reveal a significant between-groups effect for L1 alone, which I attribute to the high degree of variation. This finding is consistent with Vercellotti's (2012) lack of L1 effects in global measures of development. In addition, there was no main effect for time; nor were there effects for time interacting with L1; nor for time and CAF; and similarly, there was no significant effect for the interaction of time, L1 and CAF

measure. These results suggest that enrollment in an IEP alone is not enough to ensure global development via improvement on all CAF measures for all learners in this short time period. Again, I attribute the lack of time effects to the high degree of both inter- and intra-individual variation. Some degree of this variation can be attributed to the RSA topics, which varied across learners and levels. In addition, I would argue that learners' affect could have played a role in their CAF measures, especially if different RSA topics triggered different affective responses across learners. Such an analysis would be consistent with DST, because factors including affect, motivation, and attention are all interconnected in learners' complex adaptive L2 system.

6.1.1 Development by L1 group

Despite the lack of effects for L1 alone and for time, there was a significant interaction between the CAF measures and L1, with the Arabic learners exhibiting considerably higher fluency and slightly higher accuracy than the Chinese learners. Because this interaction was significant, I ran additional RM ANOVAs for the two L1 groups and found that the results differed by L1. The Arabic learners failed to exhibit a significant overall effect for time. Although there was a main effect for time on complexity, indicating that Arabic learners tended to improve significantly in their ability to embed multiple clauses in an AS-unit over the six data collection points, there was no main effect for either accuracy or fluency. In contrast, for the Chinese learners, there was a significant main effect for time on CAF measures. Breaking this down by measure, the ANOVA revealed that like the Arabic learners, there was no main effect for time in the accuracy measure. However, time did have a significant effect on the complexity measure, with a tendency to improve over time, just as for the Arabic learners. Finally, for fluency, there was a significant effect for time, with Chinese speakers' fluency measures increasing from Point 1 to 5 and only

decreasing at Point 6. What is most significant here is how the groups differ: while the Arabic learners did not have a significant increase in fluency over time, the Chinese learners did—a fact that I attribute to cultural background and communicative orientation.

Research Question 1 asked whether language background influenced the development of CAF over time. I had predicted that the Arabic learners would have higher initial fluency and the Chinese, higher initial accuracy and complexity. Only the former part of the prediction was true, as the learners all had comparable complexity scores not only initially, but also over time. In fact, I had predicted the Chinese learners would have higher accuracy than Arabic learners based on their experience studying English as a foreign language (EFL) in the People's Republic of China, and the emphasis placed on memorization and correctness. However, the Arabic learners were actually slightly more accurate, but it is unclear whether this is due to L1 influence (as Arabic is an inflectional language while Chinese is not) or other factors because the global accuracy measure does not distinguish between types of errors. I address this below in the discussion of specific accuracy in Section 6.2.

6.1.2 A closer look at fluency

In terms of fluency, the Arabic learners did not improve significantly over time, while the Chinese learners did. Based on cultural background and initial communicative orientation, such a result is not unexpected. The Arabic learners come from a culture where oral fluency is highly valued, both in their L1 and in L2s. Recall from Juffs and Friedline (2014) that Arabic learners comparable to those under investigation in this thesis tended to name speaking as the best way to learn new vocabulary, while Korean learners (whose educational system is similar to that of the People's Republic of China) tended to emphasize text-based methods, with some even

considering speaking the worst way to learn new words. Thus, the differential development in fluency may be due to cultural background and initial communicative orientations. The Arabic learners generally entered the ELI with relatively high fluency and therefore were able to make fewer gains, with their scores exhibiting a ceiling effect. In contrast, the Chinese learners arrived with an orientation that did not emphasize fluency. Overall, they had probably had much less previous experience speaking in English than their Arabic counterparts and therefore had more room for improvement. Although their fluency scores never quite reach the Arabic learners' mean fluency, the Chinese learners do make considerable progress over time.

Because my measure of fluency was global, it is unclear what sub-dimension of fluency the Chinese learners actually made the most gains on: breakdown fluency, speed, or repair fluency (Skehan, 2003; Tavakoli & Skehan, 2005). The WPM measure conflates these three sub-dimensions because it includes filled and unfilled pauses in the denominator and subtracts words comprising false starts, non-rhetorical repetitions, self-corrections, and hesitations from the numerator. As a result, the numeric gains in fluency may reflect improvements on any of the various sub-measures. In any case, I am concerned with what fluency gains entail with respect to the other measures of complexity and accuracy. Rod Ellis and Barkhuizen (2005) wrote, "Fluency occurs when learners prioritize meaning over form in order to get a task done. It is achieved through the use of processing strategies that enable learners to avoid or solve problems quickly" (p. 139). One such processing strategy is reliance on "chunks," formulaic phrases of memorized language. Skehan (1998) writes that during rapid communication, "We rely on such chunks to ease processing problems, using them to 'buy' processing time while other computation proceeds, enabling us to plan ahead for the content of what we are going to say, as well as the linguistic form" (p. 40). Although my coding schema did not consider the frequency

of chunks vs. language that is analyzed and constructed creatively, it is likely that enrollment in an IEP leads to frequent exposure to chunks both inside and outside the classroom, and that the fluency gains exhibited by Chinese learners could have been influenced by the incorporation of chunks. Furthermore, frequent use of hedges and chunks such as “I think” and “What I mean is” will result not only in an increase in fluency, but also in gains in complexity by subordination. Therefore, it is reasonable to argue that increased use of chunks may be behind both gains in fluency and complexity.

6.1.3 Interactions between CAF measures

Thus far, my discussion has focused on individual CAF measures and the role of L1 and time in charting CAF development. Equally interesting and worth discussing is the interaction between the individual CAF measures and whether these measures are connected growers that increase simultaneously; whether they exhibit trade-off effects (Skehan, 1998, 2003); or whether no global trend is exhibited because each learner’s interlanguage is a complex adaptive system whose development is contingent on a web of interrelated variables according to a dynamic systems theory framework.

Although Vercellotti (2012) found that CAF were connected growers for the 66 learners she investigated, a glance at my Appendix C reveals that learners did not improve in all three CAF measures linearly over time. In fact, there was no main effect for time, either alone or when considered in relation to the other variables (L1, CAF). Instead, learners tended to fluctuate across performances, sometimes with improvements in one measure, and sometimes in all three. Furthermore, some learners even regressed over time, especially in the global accuracy measure. I would argue that such variation is due both to the different RSA topics across learners, and to

more general individual differences such as communicative orientation, processing strategies, etc. In any case, controlling for the speech topic in further research would allow a more thorough investigation of the specific relationship between CAF measures.

Skehan's trade-off hypothesis (1998, 2009) is based on the notions that performance is complex and multidimensional, and that learners have limited attentional resources. Because of the competition for internal processing resources, learners cannot equally attend to all aspects of CAF simultaneously, and attention to one area often occurs at the expense of attention to others. Sometimes learners' decisions to focus on one factor are due to task demands (i.e., the RSA topic to be discussed) and varying levels of cognitive and affective involvement. At the time same, differential emphasis on the CAF measures may be due to individuals' divergent communicative orientations and a predisposition to prioritize particular areas consistently. Larsen-Freeman (2009) explains, "it is not the task characteristics alone that dictate performance; it is the interaction between the task and the task participants—in complex systems terms, the two together form a coupled system" (p. 585). Therefore, it is necessary to consider both elements in my analysis.

Skehan (1998) characterizes the primary tension in allocation of attentional resources as between meaning (operationalized through fluency) and form (complexity and accuracy). He also notes a tension between the two sub-aspects of form and characterizes two prototypical learners with divergent focuses. A learner who favors control and exhibits conservatism may be "willing to rely on less ambitious communicative aims and less ambitious form, but form which is adequately controlled and where error can be avoided" (p. 286), resulting in high accuracy and low complexity. On the other hand, a learner who emphasizes risk-taking and interlanguage change may be "willing to take on complex form and respond to challenges, but … may not have

control over such formal elements as are involved, and so will make more errors" (p. 286). Skehan writes that ideally, learners should aim for CAF growth in "productive harmony," whereby

> Progress in one would be accompanied by development in the others. Complexity-restructuring would see growth in the underlying interlanguage system matched by the development of control over the (relatively) newly acquired form, the progressive elimination of error in its use, and the integration of the form into fluent performance through its dual-coding, where appropriate, as an accessible memory-based unit. (p. 287)

This balanced path of development is good for avoiding situations there is high fluency but limited complexity and accuracy, or at the same time, very complex and/or correct language that is performed painfully slowly. In summary, the challenge facing ESL instructors is to establish principles that allow instruction to promote balanced development. This topic is addressed further in the discussion of pedagogical implications in Section 6.3.

6.1.3.1 Dynamic systems theory

This thesis reveals that no single learner exhibited linear growth in all three CAF measures over time. Although this may be due to diverse RSA topics, controlling for this variable would not necessarily result in linear growth given the presence and interaction of so many other factors. Recall that Larsen-Freeman's longitudinal 2006 study of five Chinese learners of English repeating the same task found that different learners followed different paths of development with different rates of change over time and varying ultimate attainment. Furthermore, Larsen-Freeman noted that some learners even "finished their six-month course worse off with regard to a particular CAF dimension when they had started!" (2009, p. 586). Despite the high degree of individual variation, the group averages suggested linear growth in all measures over time; therefore, Larsen-Freeman's paper serves as a warning that group averages alone may not tell the

whole story, and the trend or functional relation they illustrate may have no validity for any individual learner.

In accounting for learners' divergent paths of development and differing CAF priorities, Larsen-Freeman characterizes learners' interlanguage as a complex adaptive system (CAS) situated in a dynamic systems theory framework. What is significant about this framework is that variability is seen as an important source of information about the underlying developmental process. Larsen-Freeman (2006) explained that the emergence of CAF should not be viewed "as the unfolding of some prearranged plan, but rather as the system adapting to a changing context, in which the language resources of each individual are uniquely transformed through use" (p. 590). She characterized language performance and development as complex, nonlinear, dynamic, and socially situated processes, implying that the search for universal constraints or paths of development is ultimately misguided.

6.2 MORPHEME ACCURACY

The lack of improvement over time in global accuracy measures demanded a look at the accuracy of learners' performances through a more precise lens that distinguishes between types of errors. Although an in-depth error analysis that attributes each deviation from TL norms to either L1, universal developmental factors, or individual fossilization/idiosyncrasies would have been revealing, it was beyond the scope of this thesis. Therefore, I limited my specific accuracy analysis to correct usage of six grammatical functors whose contexts occurred frequently enough in the data to derive meaningful scores.

Research Question 2 asked the extent to which language background would influence accurate usage of six grammatical functors: nominal functors plural *–s*, definite article *the*, and indefinite article *a/an*; and verbal functors regular past *–ed*, irregular past verb forms (*went, left,* etc.), and the third person singular present tense *–s*. Based on corresponding forms and categories (or lack thereof) in Arabic and Chinese, I had predicted that the Arabic learners would be significantly more accurate on plural *–s*, *the*, regular past *–ed*, irregular past, and third person singular *–s*, with similar suppliance of indefinite article *a/an* from all learners, as this category is absent from both Arabic and Chinese. The RM ANOVA I ran did reveal a significant between-groups effect for L1, with the Arabic learners on average exhibiting higher accuracy. In addition, there was a significant main effect for time, with a tendency to improve in accuracy from Level 3 to Level 4, but not on every morpheme. However, the interactions between time and functor, as well as between time, functor, and L1 only approached but did not reach statistical significance. Again, this is likely due to the high degree of inter- and intra-individual variation.

6.2.1 Nominal functors discussion

As discussed in Section 5.2.5, the L1 differences were more evident on some functors than others and also varied by time. Let us first review the results for the nominal functors. For plural *–s*, the Arabic learners were significantly more accurate than the Chinese at both levels, which I attribute to the presence of a corresponding plural category in Arabic but not in Chinese. For the definite article *the*, one would expect to find similar results based on L1 influence, but the effects were much smaller than for plural *–s*. Perhaps this is due to the specific mismatches in definite article distribution in English vs. Arabic. In any case, the larger difference between L1 group means for *the* at Level 3 than Level 4 suggests that L1 influence plays a greater role at lower

levels of proficiency as development occurs in an instructed SLA environment. Next, for indefinite articles *a/an*, I had predicted similarly low levels of accuracy for both groups of learners, which is exactly what the data exhibit. The Chinese learners exhibited larger variation than their Arabic learners in their suppliance of this functor at both levels, perhaps due to the lack of any corresponding article category in Chinese, while Arabic has a definite article that functions similarly to that in English. In fact, Master (1997) argued that learners whose L1 has no article system (like Chinese) will have more difficulty acquiring both indefinite and definite articles than learners whose L1 features either of these categories. The lack of significant improvement in *a/an* accuracy from Level 3 to Level 4 suggests that learners are still struggling to achieve native-like control of this category, even as their global proficiency increases.

6.2.2 Verbal functors discussion

In terms of the verbal functors, I had predicted that the Arabic learners would outperform the Chinese on all three forms at both levels due to corresponding inflectional tense morphology in Arabic that is absent in Chinese. Yet the results were much more complex, revealing that L1 influence alone is not enough to determine accuracy on specific functors. For regular past *–ed*, whose accurate usage relies on syntactic processing, the Arabic learners were slightly more accurate than the Chinese at Level 3, but the group averages indicate a regression at Level 4, while the Chinese learners made significant progress and slightly outperformed the Arabic group at this second level. The large range of scores and high degree of variability indicate that any group trends are also highly susceptible to the influence of individual differences.

For the irregular past tense forms, whose accurate usage relies on lexical or memory-based processing, the Chinese learners actually performed slightly better than the Arabic group at

both levels. In Section 5.2.5, I tentatively attributed this finding to cultural background and the fact that Chinese learners may be better at memorizing and deploying forms based on the instructional and learning techniques emphasized in Chinese culture, both for learning Chinese characters and English. However, it is also likely that contextual factors such as past time adverbials (e.g., *yesterday, last month, 3 years ago*) also condition past tense usage (Bardovi-Harling, 2000; N. Ellis, 2006; VanPatten, 2007). First language learners of English acquire past tense marking long before temporal adverbs, while L2 English learners establish temporal reference first with adverbials and only later with verbal morphology (Bardovi-Harling, 1992). Regardless of L1, the presence of such time adverbials may inhibit past tense suppliance for some learners, as any referential information encoded by the past tense is redundant when the time frame is already given elsewhere. Yet for other learners, these adverbs may function as a *cue* to supply the accurate past tense form (be it regular *–ed* or irregular forms). For this reason, future research should look more closely at the context of each past tense suppliance or omission.

Lastly, if both regular and irregular past are considered together, the Arabic learners' regression on regular past *–ed* from Level 3 to Level 4 is accompanied by a significant improvement in irregular past suppliance, suggesting that trade-offs within accuracy may be evident if learners focus on lexical processing at the cost of syntactic processing. As mentioned earlier, is also possible that the Arabic learners' representation of regular past *–ed* is undergoing restructuring at Level 4, occupying the nadir of a U-shaped learning curve (R. Ellis, 1987). Basing their research on Ullman's (2001) model of declarative and procedural knowledge, Morgan-Short, Finger, Grey and Ullman (2012) argued that it takes time for explicit L2 declarative knowledge to undergo restructuring and become implicit procedural knowledge, characterized by native-like processing. In this case, I would argue that the Arabic learners'

representation of the regular past *–ed* is undergoing such restructuring, and not enough time has passed over the course of observations to allow "the consolidation of knowledge in declarative and procedural memory, on which L2 grammar learning appears to depend" (Morgan-Short et al., 2012, p. 1). Hence, there is a decline in accuracy from Level 3 to Level 4. Yet this hypothesis can only be investigated if learners' specific accuracy at Level 5 is also considered, as they (hopefully) continue to improve in proficiency.

The results for the third person singular *–s* morpheme reveal a high degree of inter-individual variation at both levels, but with a general tendency to improve over time. As discussed earlier, the characteristic undersuppliance of this morpheme may be due to the fact that it will always encode redundant information, as English requires the inclusion of subjects before finite verbs. Furthermore, its position in a coda is not particularly salient and makes this morpheme highly susceptible to deletion. A pilot study preceding this thesis was based on the hypothesis that phonological factors could explain Arabic learners' lower suppliance of functors that are realized on a complex coda than on an open or simplex coda. However, the results of this study were not statistically significant. The conclusion is that phonological factors alone cannot determine suppliance, because if this were the case, then one would expect the same accuracy on plural *–s* and third person singular *–s*. Therefore, it is a combination of factors, including but not limited to context and redundancy, that determine accurate usage of this morpheme.

On a final note, based on the larger difference between the two groups' mean accuracy on *the* at Level 3 than at Level 4, I had reasoned that L1 influence was stronger at lower levels of proficiency. If this were the case, then Arabic learners would be expected to be more accurate on third sing. *–s* than their Chinese counterparts at Level 3, and for the scores to "even out" at Level 4. However, there was similar undersuppliance for all learners at Level 3, while at Level 4, more

Arabic learners than Chinese learners supplied it at all. Because third person singular *-s* is the last morpheme to emerge, I would argue that the L1 influence also emerges later, as acquisition of this morpheme occurs later than the others.

6.2.3 The reliability of the grammatical functor analysis

Although L1 did play a significant role in suppliance of the grammatical functors under investigation here, the two implicational scales revealed that L1 is not the only conditioning factor in specific accuracy. If this were the case, then I would expect learners of different L1s to exhibit divergent hierarchies and for all Arabic learners to outperform the Chinese on all functors but indefinite *a/an*, which was clearly not the case. Furthermore, if L1 alone determined suppliance, then learners from the two L1 groups would be expected to cluster together on the implicational tables, with all Arabic learners located on the lower half, and with more total asterisks than the Chinese group. Instead, at both levels the learners were interspersed throughout the tables, just as Spinner (2007, 2011) found in her investigation of Chinese, Korean, Spanish and Arabic learners of English. Similarly, although the functors under investigation exhibited a crude order of emergence, this order did not apply to all learners and could not be used to predict individual results, indicated by the non-significant coefficients of reproducibility at both levels.

Similarly, the order of emergence is much less rigid than suggested by Krashen (1977) and Goldschneider and DeKeyser (2001), with significant inter-individual variation, and intra-individual variation across levels. For example, recall that the Level 3 rank order was *the* > plural *–s* > *a/an* > reg. past *–ed* > irreg. past > third singular *–s.* Yet among my pool of 30 learners, *not a single learner's* accuracy scores matched this hierarchy. At Level 4, the rank order was *the* > plural *–s* > irreg. past > *a/an* > reg. past *–ed* > third singular *–s*. Again, *not a single learner's*

scores correspond to this hierarchy. Many of the deviations consist of learners scoring higher on plural *–s* than *the* (e.g., Level 3: A 12, A 45, A 157, A 163, A 241, A 404, A 481, A 530, C 126, C 282, C 298, C 537, and C 631; at Level 4: A11, A 129, A 157, A 481, C 126, C 177, C 298, and C 520). There are also a number of learners who performed better, for example, on irregular past tense forms than the article *a/an* at Level 3 (e.g., A 25, A 29, A 45, A 157, A 159, C 126, C 177, C 270, C 282, C 456, C 631, and C 914). These trends suggest the significant role of both L1 and individual idiosyncrasies. They also remind us of the statistical pitfall highlighted by Larsen-Freeman: the trend illustrated by averages of aggregated data may have no validity for any one individual. Despite Goldschneider and DeKeyser's (2001) in-depth meta-analysis, the data in this thesis do not meet the authors' predictions—thus, single explanations relying only on "salience" are insufficient to account for the trends observed.

6.3 GENERAL DISCUSSION OF IMPLICATIONS

In this section, I would like to discuss the implications of four aspects of this research: first, how evidence of L1 influence depends on the grain size of analysis (6.3.1); second, other individual differences and paths of development (6.3.2); third, pedagogical implications based on these individual differences (6.3.3); and finally, implications for further research employing CAF as dependent variables (6.3.4).

6.3.1 Language background and cultural influence

Recall that on the CAF measures, there was no global between-L1 groups effect, with the only significant difference between the groups occurring on the fluency measure, with the Arabic learners consistently scoring higher than the Chinese group, but only the Chinese learners improving significantly over time. I credit the different fluency performances to cultural background, as oral proficiency in a foreign language is emphasized in Arabic culture but largely ignored in Chinese culture due to its exclusion on high stakes standardized exams and limited opportunities for oral practice with native speakers of English (Chen et al., 2005).

On the other hand, there was a reliable between-groups effect for the grammatical functor analysis. This finding reveals that the grain size of the analysis can have a significant effect on the results, where L1 effects are only evident in a microanalysis of learners' errors. The implicational scales, however, indicate that L1 alone is not enough to predict learners' accuracy on six grammatical functors. For example, some Chinese learners placed high, while other Arabic learners had low scores, suggesting that L1 is only one of myriad factors that can determine specific accuracy. Other factors that influence not only specific accuracy but also the global CAF measures include the RSA topic, individual learners' communicative orientation (i.e., emphasis on meaning vs. form, risk-taking vs. control), personality, affect, age, social distance, gender, and level of motivation, among others (Gass & Selinker, 2008).

6.3.2 Other individual differences: Meaning vs. form orientations

Although it is difficult to operationalize all of the individual differences outlined above, it is possible to consider communicative orientations and emphasis on meaning vs. form via frameworks that highlight different paths of development.

Skehan (1998) argues that the differences between learners with varying emphases on CAF can be related to their individual differences in processing strategies. Some learners are more analytic, which might result in higher complexity at the cost of fluency and accuracy, while others are predisposed towards memory, and the greater reliance on chunks could lead to higher complexity, accuracy, and fluency. Although Skehan suggested balancing complexity restructuring, accuracy, and fluency in general so learners don't prioritize one aspect over the other, balance over time for *particular* learners is also critical. Skehan (1998) writes, "Learners who might prefer to emphasize fluency (say) would then need to be treated slightly differently from learners who prioritize form, either for accuracy or complexity. Learners, that is, may prefer to do what comes naturally to them, even though this may have unfortunate consequences for longer-term development." (p. 289). As a result, it becomes the teacher's task to address individual learners' preferences and difficulties so that, for example, the fluency-oriented learners can address their problems with form.

6.3.3 Pedagogical implications

Although this thesis found that learners neither exhibited universal trade-off effects nor simultaneous growth over time in their CAF scores, the findings show that improvement did not occur globally or linearly for all learners on all measures. Instead, for individual learners,

improvement in one areas sometimes occurred at the expense of improvement elsewhere. However, in Skehan's (1998) words,

> [T]here are encouraging signs that task characteristics predispose learners to channel their attention in predictable ways, such as clear task macrostructure toward accuracy, the need to impose order on ideas towards complexity, and so on. Obviously these interpretations are post hoc and need to be validated through further research. But they are suggestive, and imply that, if such results can be replicated, tasks may be chosen and implemented *so that particular pedagogic outcomes are achieved.* (p. 112, emphasis original)

The task design and implementation to which Skehan refers may either apply to all learners or to individuals and their unique preferences toward focusing on one aspect of the CAF triad.

In terms of curricular design and task characteristics, Skehan (2003) suggested that structured tasks (i.e., with a clear timeline or macro-structure) encourage greater fluency in learner performances, as well as a tendency toward greater accuracy (Foster & Skehan, 1996; Skehan & Foster, 1997, 1999). When structured tasks are combined with delayed processing conditions, the greatest accuracy is achieved. Next, tasks that rely on familiar information will result in greater fluency and greater accuracy. With respect to implications for individual learners, instructors could implement individualized feedback that highlights learners' strengths and weaknesses and suggests where to focus their limited attentional resources in future performances.

Another pedagogical implication is the value of task repetition. In this research, each time learners performed an RSA, the topic varied but the task conditions were consistent. Because the learners received individualized feedback on prior performances before undertaking consequent ones, perhaps CAF growth could be due to following their instructors' feedback. But it is also possible that learners grew more familiar with the task demands and were able to transfer this knowledge to later speeches. Rod Ellis (2009) argues that learning has occurred when learners

demonstrate that they can transfer what they have learned to a new task. In fact, as discussed earlier, Ahmadian (2011) found that learners who repeated a task 11 times every two weeks outperformed a control group on a new task in complexity and fluency, but not accuracy. However, this type of "transfer" cannot necessarily be expected to occur among the learners studied in this thesis, as the conceptual demands of each RSA depend on the topic, which varied across learners. Furthermore, Larsen-Freeman (2009) reminded readers that from a dynamic systems perspective, the lack of immediate evidence of task repetition assisting learning does not imply that learning has not occurred. Instead, it is plausible that there exists a nonlinear relationship between a learner repeating a task and that same learner showing improvement from a TL perspective (p. 584). That is to say, the benefits that come from repeating a task may not be immediately evident; for this reason, researchers must think longitudinally and non-linearly.

6.3.4 Implications for the measurements of CAF

This research operationalized language performance through the three CAF constructs. Complexity, fluency, and global accuracy were assessed via general, rather than specific measures, and specific morphosyntactic accuracy was measured by target like usage of six grammatical functors. The effectiveness of each construct is discussed below.

Complexity was measured via syntactic complexity by subordination, calculated by dividing the average number of clauses per AS unit per RSA. This measure was an effective operationalization of development, as most learners employed more subordinate clauses in their RSAs over time. However, this measure of complexity will only increase as the number of embedded clauses grows, and will not reflect language that is more complex because of phrasal elaboration, expected to occur at higher levels of proficiency. Norris and Ortega (2009)

suggested also measuring complexity by the mean number of words per clause and per AS-unit, as Vercellotti (2012) did. Although I originally coded and calculated each learner's speech according to these additional measures of global complexity (words per AS-unit) or phrasal elaboration (words per clause), in the end these measures were not used because they demanded more complex statistical analysis than that employed in this thesis. This was a methodological, not a theoretical issue. However, I would suggest that future research also consider these measures, as they capture different kinds of syntactic complexity than the subordination measure. In addition, lexical variety is an important measure of language development, but the differing RSA topics among learners introduced variation that obviated any type of lexical variety analysis.

The global accuracy measure of the average number of error-free clauses divided by total clauses could have also been supplemented by a ratio of error-free AS-units. However, I decided that this latter measure was not well suited to lower proficiency learners, as many struggle to produce a single error-free AS-unit, especially as the length of the AS-unit increases. Vercellotti (2012) calculated both clausal and AS-unit accuracy and found that these measures had significant between- and within-individual correlations, another reason I did not calculate this latter measure. However, this global clausal accuracy measure does not reflect the type or source of errors, the length of the clause, nor the number of errors in a clause. Hence, a learner who produces a relatively long clause with one error will be faulted the same as a learner whose shorter clause has, say, five errors. For this reason, perhaps a measure such as errors per 100 words could control for length of the unit and number of errors per unit. Yet because such a unit lacks both psychological and linguistic reality, it has been largely abandoned by the field (Vercellotti, 2012, p. 170).

The measure of specific accuracy on six grammatical functors was a valid way to capture L1 influence where the global accuracy measure fell short. Giving half-credit for misformations (à la SOC, suppliance in obligatory contexts) was an effective way to capture partial accuracy. For example, learner C 127's utterance "this is **a** important holiday in China" contains a misformed indefinite article, indicating that the learner knew which article to use (among definite, indefinite, and Ø) and simply supplied the wrong allomorph. Including oversuppliance in the denominator (à la TLU, target-like usage) was an effective way to capture overgeneralizations such as "in **the** life", which were present in many learners' performances and tend to be very common for learners in an instructed SLA environment (Pica, 1983). However, one problem with this measure is that I included self-corrections in the misformations figure. As Rod Ellis and Barkhuizen (2005) write, "The number of *self-corrections* does not provide a measure of how accurately a learner uses the L2 but rather indicates the extent to which the learner is oriented toward accuracy" (p. 149-50). Thus, although self-corrections are a valuable measure and may reveal much about a learner's orientation, it is the topic of another research agenda. For future research, I would modify the specific accuracy measure to assign full credit for self-corrected grammatical functor suppliance, not counting any preceding formulations.

Finally, the measure of fluency in average words per minute (WPM) is a good indicator of average rate of speech, but it fails to individuate breakdown and repair fluency (Tavakoli & Skehan, 2005). In order to quantitatively measure these two sub-dimensions of fluency, it is necessary to operationalize breakdown fluency via the number of pauses and average pause length, and repair fluency by measuring false starts, reformulations, and replacements. Instead, my fluency measure conflated all three sub-dimensions. In addition, perhaps a speed fluency

measure would be more accurate if it measured not words but syllables per minute, as English words may contain a considerable range of syllables.

On a final note, my fluency via WPM measure did not consider the content of the words uttered and the adequacy of the message. A learner may have a high rate of fluency but may not effectively communicate his message due to frequent errors such as lexical choice and “talking in circles.” Therefore, future researchers are urged to also consider communicative adequacy as another aspect of development.

7.0 CONCLUSION

7.1 SUMMARY AND DISCUSSION

This longitudinal research illustrated a lack of between-group L1 effects in the global CAF measures, but a significant interaction between L1 and CAF, with Arabic learners exhibiting higher fluency and slightly higher accuracy than the Chinese students. In addition, there was a significant between-L1 groups effect for accuracy on six grammatical functors.

The results of the inferential statistics indicate that L1 effects are more evident in specific measures of accuracy than general ones. This suggests that future research should employ both local and global measures of accuracy, as well as complexity and fluency, in order to better understand how language background effects manifest themselves in L2 performance.

On a related note, language background alone could neither account for nor predict specific accuracy, as learners with the same L1 did not cluster together on the implicational scales. This implies that L1 may not be the best grouping factor when considering language performance, as a few of the Chinese learners exhibited highly accuracy plural usage, suggesting that they are acting more like the Arabic participants! Similarly, as mentioned earlier, aggregating data and comparing group means does not reflect individual tendencies. As Skehan

(2009) emphasizes, it is necessary to run intra-individual in addition to group correlations, as group averages may illustrate a trade-off or growth trend that does not apply to individuals.

Another key issue illustrated by this thesis is the high level of variation between and among individual learners. There is much debate in the field of SLA about the significance of such variation. Larsen-Freeman (2009) argued that difference and variation need to move to the center of SLA research, ascribing a dynamic quality to individual differences. However, Pallotti (2009) disagreed with Larsen-Freeman that variation should occupy the forefront of CAF research. He described the "the necessary variation fallacy" as researchers' tendency to seek to identify measures of language performance that most clearly show variance among subjects over time and across tasks, correlating with other equally varying proficiency measures (p. 590). However, in Pallotti's words, "a measure can be scientifically valid and informative even if it does not show any difference among groups of subjects" (p. 590). In his opinion, research should not only be concerned with differences and variation, but also with constants and similarities. This research tried to find a balance between both and found that although some constants do exist (e.g., that Arabic learners generally have a higher rate of fluency than Chinese students), they are subject to considerable and significant variation, which always returns to the importance of differences between individuals.

7.2 LIMITATIONS

This study is not without its limitations, including the small sample size of N = 30 participants and less than ideal data collection instruments. Because the data come from the ELI Online Database, the students were enrolled over different semesters and therefore had different instructors and speaking topics, both factors that might unnecessarily introduce additional variation into CAF scores. In terms of how RSA topic might affect performance, for example, Eisenstein and Starbuck (1989) found that learners focused more on meaning (with consequent drops in accuracy) when discussing a topic of personal interest, while the same learners focused more on form when given a topic in which they were not personally invested. Shirai (1992) noted greater L1 influence in L2 production when discussing a topic strongly connected to the L1 and concluded, "if the L1 conceptual structure is activated, L2 performance will be influenced by L1" (p. 110). Therefore, future research should consider students enrolled over the same semester in order to control for RSA topic. With this source of between-individual variation attenuated, it would also be possible to also measure the development of complexity via lexical variety.

In addition, this research only investigated six two-minute oral monologues per learner (12 minutes total) and did not directly explore processing strategies, comprehension, or communication strategies. One advantage to the data, however, is that the learners were not aware that their production would be analyzed for research and the data came from naturally occurring classroom exercises.

7.3 FUTURE RESEARCH

Future research on the role of language background in CAF development has a bright future. Critical to future research is the need to employ more advanced statistical analyses. This research relied on mixed RM ANOVAs and implicational scales, but additional correlations could reveal more about trade-off effects. Although I also considered employing hierarchical linear modeling (HLM) as Vercellotti (2012) did, this type of multilevel model assumes a rectilinear relationship between variables. Although HLM can also model non-linear change trajectories, the advantages offered by HLM (controlling for attrition, missing data, different spacing between observations, and predictor variables) were not necessary for my data set, and two RM mixed ANOVAs and implicational scales were sufficient to explore my research questions.

Future research should also consider a larger sample of learners from additional language backgrounds and over a longer period of time. The results of the regular past *–ed* morpheme measures suggest that learners' knowledge may be undergoing restructuring, as the Arabic learners regressed from Level 3 to 4. However, this trend does not mean they are not learning; rather, their syntactic knowledge is undergoing restructuring, and it takes time (often months or years) for this representation to become consolidated so that it can be accessed and applied in a native-like way (Morgan-Short et al., 2012). In order to confirm this explanation, it would be necessary to look at their performances at Level 5. With respect to language background, it would be interesting to consider learners whose L1 is closer to English (e.g., Romance and Germanic languages) to see whether their specific accuracy scores change at the same rates as the Arabic and Chinese learners considered here. In addition, further research is required to

confirm the effects of cultural background and the emphasis on L2 oral proficiency in a variety of EFL contexts.

As far as accuracy goes, this research illustrates how L1 effects may only be evident on a specific, not global level. Future research should look closer at the type of error, be it lexical, syntactic, or morphological, and see what percentage of errors comprise each category for different language groups. This could also have pedagogical implications, as instructors could provide more individualized feedback if it was found, for example, that Arabic learners make more lexical errors while Chinese learners make more morphological ones.

This research has also illustrated the dangers of statistical analysis and the potential of group scores to misrepresent individual tendencies. For this reason, applied linguists and researchers in the social sciences in general should always confirm that group scores have individual validity (Skehan, 2009). Within the realm of CAF research, this thesis has also illustrated that general CAF measures fail to paint a full picture of developmental trends and the role of L1. For this reason, it is critical to adopt both global and specific measures in order to better understand the complex, dynamic process that is SLA.

APPENDIX A

LEARNERS' DEMOGRAPHIC INFORMATION

Table 22. Demographic information

Learner ID	Age at data collection	Gender	Level 3 semester of enrollment	Level 4 semester of enrollment	MTELP conversion score	Listening score	Writing Score
A 11	18	M	summer 2006	fall 2006	40	13	2
A 12	18	M	summer 2006	fall 2006	39	14	1.8
A 25	25	M	summer 2006	fall 2006	28	5	1
A 29	20	M	summer 2006	fall 2006	38	7	1.8
A 30	19	M	spring 2006	summer 2006	42	17	2.5
A 45	18	M	summer 2006	fall 2006	32	6	1
A 65	23	M	spring 2006	summer 2006	51	16	2.7
A 129	18	M	spring 2006	summer 2006	36	12	2.5
A 157	23	M	summer 2006	fall 2006	48	9	3
A 159	22	M	summer 2006	fall 2006	48	15	3
A 163	26	M	summer 2006	fall 2006	60	11	2
A 241	24	M	summer 2007	fall 2007	41	12	3
A 404	19	M	spring 2007	summer 2007	41	12	1.7
A 481	27	M	summer 2007	fall 2007	42	10	2
A 530	22	M	fall 2007	spring 2008	45	9	3.3
C 126	37	F	summer 2006	fall 2006	25	8	2.5
C 127	29	F	spring 2006	summer 2006	47	12	2.7
C 177	23	F	summer 2006	fall 2006	51	13	3
C 270	22	M	fall 2006	spring 2007	35	12	1.5
C 282	31	M	fall 2006	spring 2007	74	16	4
C 298	26	F	fall 2006	spring 2007	57	16	4
C 301	29	F	fall 2006	spring 2007	55	16	4
C 456	23	F	summer 2007	fall 2007	45	14	2
C 520	27	M	fall 2007	spring 2008	48	10	2.8
C 537	29	F	fall 2007	spring 2008	46	10	1

C 611	26	M	spring 2008	summer 2008	43	12	3.6
C 631	25	F	summer 2008	fall 2008	42	14	2.8
C 633	28	F	summer 2008	fall 2008	47	13	3
C 914	34	F	summer 2009	fall 2009	45	12	2.8
C 988	19	M	fall 2009	spring 2010	38	7	2

APPENDIX B

RSA TOPICS AND PROMPTS

B.1 LEARNERS' RSA TOPICS

Table 23. RSA topics by learner

Learner ID	Level 3 RSA #1	Level 3 RSA #2	Level 3 RSA #3	Level 4 RSA #1	Level 4 RSA #2	Level 4 RSA #3
A 11	best friend	funny or scary experience	favorite holiday	Pets – B	important person in my past	biggest problem in my country
A 12	best friend	funny or scary experience	favorite holiday	Pets – B	important person in my past	biggest problem in my country
A 25	best friend	funny or scary experience	favorite holiday	Pets – B	important person in my past	biggest problem in my country
A 29	my country	funny or scary experience	favorite holiday	Pets – B	important person in my past	biggest problem in my country
A 30	my background	important event in my country-a	a place you like	favorite place	funny or scary experience	favorite holiday
A 45	best friend	funny or scary experience	favorite holiday	Pets – B	important person in my past	biggest problem in my country
A 65	sports	upcoming vacation	important event in my country -a	favorite place	funny or scary experience	favorite holiday
A 129	sports	upcoming vacation	important event in my	Pets – A	funny or scary	favorite holiday

			country-a		experience	
A 157	my country	funny or scary experience	favorite holiday	pets – B	important person in my past	biggest problem in my country
A 159	best friend	funny or scary experience	favorite holiday	pets – B	important person in my past	biggest problem in my country
A 163	best friend	funny or scary experience	favorite holiday	pets – B	important person in my past	biggest problem in my country
A 241	my city	first school	most important things	free time	significant event	cultural differences
A 404	shopping for food	can't do here	custom in your country	my city	first school	most important things
A 481	my city	first school	most important things	free time	significant event	cultural differences
A 530	free time	significant event	cultural differences	important event in my country – B	important person in my country	famous place
C 126	best friend	funny or scary experience	my favorite holiday	pets – B	important person in my past	biggest problem in my country
C 127	sports	upcoming vacation	important event in my country – A	pets – A	funny or scary experience	favorite holiday
C 177	my country	funny or scary experience	my favorite holiday	pets – B	important person in my past	biggest problem in my country
C 270	pets – B	important person in my past	biggest problem in my country	shopping for food	can't do here	custom in your country
C 282	pets – B	important person in my past	biggest problem in my country	shopping for food	can't do here	custom in your country
C 298	pets – B	important person in my past	biggest problem in my country	shopping for food	can't do here	custom in your country
C 301	pets – B	important person in my past	biggest problem in my country	shopping for food	can't do here	custom in your country
C 456	my city	first school	most important things	free time	significant event	cultural differences

C 520	free time	significant event	cultural differences	important event in my country – B	important person in my country	famous place
C 537	free time	significant event	cultural differences	important event in my country – B	important person in my country	famous place
C 611*	important event in my country – B	important person in my country	famous place	learning English	----	foreign language
C 631	my background	country change past 50 years	next vacation	greatest accomplish ment	local customs	problem
C 633	my background	country change past 50 years	next vacation	greatest accomplish ment	complaint	problem
C 914	life pre-ELI	first day in Pittsburgh	my favorite holiday	a trip	university in my country	strategies to improve English.
C 988	childhood	Traveling in my country	confusing situation	job	vacation spot	renting

NB: Because learner 611's second Level 4 RSA was absent from the database, this participant was removed from the CAF analysis.

B.2 RSA PROMPTS BY TOPIC

Table 24. RSA topics and prompts

Topic	Prompt
best friend	Talk about your best friend.
my country	Describe your country or an interesting place in your country.
funny or scary experience	Talk about a funny or scary experience that you had.
favorite holiday	Talk about your favorite holiday.
my background	Talk about your background.
important event in my country – A	Talk about an important event that happened in the past in your country.
a place you like	Talk about a place that you really like. Describe it and tell why you like it.
sports	What sports do you enjoy?

upcoming vacation	Where do you want to go on vacation?
next vacation	Describe your next vacation.
my city	What city do you come from? Describe your city. Are there some places in the city that are dangerous? Describe those parts of the city. Are there some places that are safe? Describe those places.
first school	Describe your experience in your very first school. How old were you? How many other children were in your class? Was it exciting or frightening or both? What did you do everyday in school?
most important things	What do you think are the most important things in life? How was your opinion changed over the years? Why has your opinion changed or not changed?
shopping for food	Is shopping for food in your country the same as in the US? Explain how it is different and how it is the same.
can't do here	Describe something that you liked to do when you were in your country but that you can't do here. Where did you do this? Why did you like it? How did it make you feel?
custom in your country	Choose a custom (baby's birth, wedding, funeral, entry to adulthood, etc.) in your country. Describe what is done for this custom and why.
country change past 50 years	Talk about how your country has changed in the past 50 years.
complaint	Describe a situation when you had a complaint about something, a product you bought, a problem with your apartment, a meal in a restaurant. What was your complaint? Why did you have the complaint? What did you do about the complaint? How was the situation resolved?
favorite place	Talk about one of your favorite places.
Pets – A	How do people in your country feel about pets?
Pets – B	How do you feel about pets? Do many people have pets in your country? How are they treated, in general?
important person in my past	Talk about a person who was very important to you in the past. Who was this person? Why was this person important to you?
biggest problem in my country	What is the biggest problem your country is facing today? How would you change it?
free time	Describe the kinds of things you like to do when you have free time.
significant event	Describe a significant event in your life. When did it happen? How did it happen? How did it change you?
cultural differences	Think about one aspect of culture at home and in the United States, for example, roles of men and women, working, friendships, social events with friends. In what ways is the culture here the same and/or different from your culture at home?
important event in my country – B	Describe an important event in your country's history. Do you think this event was important? Why or why not? Give two to three reasons.
important person in my country	Talk about a person who was very important in the history of your country. Who was this person? Why was this person important? Give two or three reasons.
famous place	Describe a famous place you once visited. Where was it, and what was it

	like? Why is this place famous or important? Give two or three reasons.
learning English	Describe your experience learning English in your country. Was it easy or hard to learn? What kinds of things do you do to learn English? Give two or three examples.
foreign language	Describe what you think is necessary to learn a foreign language. What are some things that interfere with successful language learning? What are things that you do that are good for your language learning and what are things that are not good for your language learning? How can you improve?
greatest accomplishment	Talk about your greatest accomplishment in life. What did you do? Why was this your greatest accomplishment? What characteristics do you have that helped you reach this accomplishment?
local customs	Talk about some local customs that you think visitors should know if they visit your country.
problem	Describe a problem that you or someone you know or knew had. What suggestions could you make for solving this problem?
life pre-ELI	Talk about your life previous to coming to the ELI. Where did you live? Whom did you live with? What were you doing? Why did you decide to study English and why did you decide to come to the ELI?
first day in Pittsburgh	What was your first day in Pittsburgh like? Describe it.
a trip	Describe a recent trip.
university in my country	Talk about going to university in your country.
strategies to improve English	Talk about strategies you use to improve your English.
childhood	Describe your best friend from childhood. How did you meet? What qualities help describe your friend? What did you use to do together?
travelling in my country	What advice would you give someone who wanted to travel in your country? Where should they go?
confusing situation	Describe a confusing situation.
job	Describe a job that you would love to have. What are the expectations for this job? What are things that you would love about this job?
vacation spot	Talk about your ideal vacation spot. What will you do there? What are some things you will miss from home?
renting	Talk about renting an apartment, either in your country or in Pittsburgh.

APPENDIX C

CAF SCORES PER OBSERVATION

NB: Each learner occupies two rows: the first row contains CAF scores for the first three observation points, while the second row contains scores for the latter three observation points. C corresponds to syntactic complexity by subordination, A to global accuracy in error-free clauses, and F to fluency in words per minute.

Table 25. Individuals' CAF scores per observation

Learner ID	C	A	F	C	A	F	C	A	F
A 11	1.154	0.467	51.282	1.556	0.655	96.111	1.353	0.478	77.143
	1.389	0.720	81.207	1.421	0.667	93.333	2.222	0.600	84.915
A 12	1.105	0.636	70.862	1.105	0.667	76.410	1.000	0.333	69.558
	1.095	0.818	76.316	1.391	0.656	88.448	1.438	0.783	73.274
A 25	1.000	0.462	34.500	1.000	0.444	28.889	1.308	0.353	44.211
	1.417	0.412	48.500	1.900	0.316	56.500	1.875	0.600	50.609
A 29	1.588	0.667	76.410	1.222	0.409	64.068	1.500	0.714	70.769
	1.833	0.208	81.770	1.714	0.583	82.373	1.857	0.385	80.500
A 30	1.852	0.520	138.305	1.645	0.510	128.967	1.417	0.676	127.438
	2.048	0.512	110.769	2.056	0.676	130.678	2.000	0.500	132.653
A 45	1.083	0.769	40.588	1.385	0.611	56.379	1.400	0.500	41.379
	1.636	0.722	51.000	2.800	0.643	50.000	2.000	0.500	66.429
A 65	1.667	0.689	108.500	1.667	0.680	87.273	1.600	0.583	85.714
	1.650	0.697	90.252	1.850	0.622	97.949	1.583	0.684	68.348
A 129	1.647	0.750	93.418	2.154	0.821	71.455	1.667	0.400	78.058
	1.500	0.750	81.026	1.857	0.577	92.308	1.588	0.704	86.441
A 157	1.182	0.462	58.421	1.500	0.444	72.432	1.500	0.389	78.103
	1.538	0.450	70.526	1.846	0.417	72.632	3.600	0.722	60.000

A 159	1.091	0.417	38.919	1.273	0.429	41.538	1.857	0.769	50.270
	1.600	0.688	46.271	1.875	0.800	57.458	1.667	0.200	50.769
A 163	1.789	0.529	85.424	1.333	0.417	86.379	1.769	0.478	90.256
	1.600	0.438	64.034	2.375	0.605	102.203	2.214	0.452	107.797
A 241	1.313	0.619	93.051	1.500	0.333	55.424	1.813	0.448	73.000
	2.071	0.724	64.000	2.000	0.625	60.612	1.923	0.520	86.218
A 404	1.154	0.200	68.136	1.571	0.500	69.310	1.200	0.417	51.282
	1.267	0.579	77.436	1.400	0.619	77.228	2.273	0.720	63.214
A 481	1.200	0.500	55.862	1.923	0.200	67.179	1.750	0.190	62.521
	2.875	0.565	52.881	1.600	0.375	50.172	1.818	0.400	62.500
A 530	2.222	0.550	55.932	1.417	0.647	54.643	1.357	0.571	85.500
	2.091	0.478	67.000	1.600	0.688	65.172	1.231	0.625	57.895
C 126	1.308	0.471	33.846	1.308	0.588	40.345	1.091	0.250	45.391
	1.286	0.333	30.275	2.167	0.615	31.525	1.333	0.333	39.273
C 127	1.273	0.357	50.092	1.333	0.625	41.157	1.889	0.412	42.720
	1.471	0.480	61.565	1.875	0.400	81.000	1.818	0.500	67.304
C 177	1.375	0.727	77.647	1.850	0.649	92.000	1.750	0.571	75.000
	1.526	0.621	88.000	1.750	0.476	63.000	2.714	0.632	57.931
C 270	1.154	0.533	48.305	2.111	0.263	58.983	1.571	0.591	63.621
	1.214	0.294	56.975	2.231	0.655	82.957	1.667	0.500	58.983
C 282	1.308	0.471	57.391	1.462	0.526	58.487	1.636	0.444	51.795
	1.600	0.625	55.000	2.000	0.708	69.500	1.857	0.692	53.500
C 298	1.118	0.526	56.923	1.364	0.400	50.556	1.900	0.579	58.889
	1.500	0.200	69.391	1.375	0.545	79.322	1.667	0.400	67.000
C 301	1.533	0.652	60.000	2.000	0.636	61.026	2.231	0.621	84.407
	1.769	0.609	85.424	2.533	0.868	98.000	1.909	0.619	70.862
C 456	1.333	0.500	62.368	1.444	0.462	93.600	2.214	0.548	93.333
	2.538	0.727	95.106	2.000	0.556	64.752	1.833	0.455	79.500
C 520	1.667	0.750	49.231	1.556	0.571	38.609	1.375	0.545	45.500
	1.444	0.769	46.667	1.300	0.692	52.308	1.909	0.905	64.068
C 537	1.500	0.444	24.000	1.667	0.400	30.000	1.600	0.375	33.913
	1.200	0.000	35.556	1.556	0.214	44.615	1.583	0.579	58.435
C 631	1.357	0.632	53.846	1.714	0.333	41.416	1.556	0.429	39.661
	2.444	0.455	51.795	2.571	0.444	47.500	2.000	0.625	84.500
C 633	1.308	0.412	48.500	1.400	0.571	45.254	1.500	0.417	59.492
	1.500	0.524	51.500	1.778	0.313	54.407	1.900	0.526	66.610
C 914	1.222	0.364	42.564	1.188	0.579	49.500	1.692	0.273	57.966
	1.300	0.538	45.254	1.533	0.522	70.862	3.000	0.583	51.000
C 988	1.455	0.313	47.767	1.692	0.318	66.154	1.286	0.042	48.205
	1.833	0.318	58.462	1.727	0.526	65.500	1.636	0.500	66.111

APPENDIX D

GRAMMATICAL FUNCTOR SCORES

PL = plural –*s*; -ED = regular past –*ed*; IRP = irregular past; 3S = third person singular present –*s*; N/A = no occurrences or contexts

Table 26. Individuals' grammatical functor scores by level

	Level 3						Level 4					
Learner ID	PL	A/AN	THE	-ED	IRP	3S	PL	A/AN	THE	-ED	IRP	3S
11	0.45	0.73	0.70	0.67	0.63	0.25	0.95	0.92	0.47	0.67	0.62	0.00
12	0.94	0.64	0.92	0.25	0.50	0.33	0.92	0.67	0.95	0.00	0.60	1.00
25	0.67	0.25	0.67	0.50	0.58	0.00	0.59	0.73	0.75	0.38	0.00	0.38
29	0.56	0.20	0.63	0.80	0.71	N/A	0.44	0.76	0.67	0.67	0.64	0.00
30	0.78	0.56	0.95	0.64	0.38	0.00	0.74	0.63	0.75	0.57	0.89	0.00
45	1.00	0.83	0.67	N/A	0.88	0.00	0.60	0.80	0.63	0.00	0.50	0.00
65	0.68	0.45	0.69	1.00	0.00	0.38	0.71	0.20	0.86	0.61	0.89	1.00
129	0.60	0.63	0.87	0.50	0.25	N/A	0.86	0.72	0.77	0.80	0.43	1.00
157	0.83	0.75	0.71	0.75	0.90	0.25	0.96	0.40	0.62	0.38	0.81	1.00
159	0.85	0.50	1.00	0.58	0.67	0.00	0.79	0.67	0.81	0.75	0.43	1.00
163	0.92	0.50	0.69	0.00	0.31	0.00	0.63	0.80	0.83	0.60	0.44	0.00
241	0.71	0.84	0.67	0.71	0.75	N/A	0.60	0.50	0.62	0.67	0.83	0.00
404	0.63	0.62	0.48	1.00	N/A	0.00	0.73	0.50	0.83	0.25	0.78	1.00
481	0.32	0.80	0.25	0.00	0.29	0.00	0.60	0.40	0.50	0.50	0.50	0.00
530	0.81	0.43	0.71	1.00	0.00	N/A	0.90	0.57	0.73	0.77	0.75	1.00
126	0.43	0.60	0.20	0.80	0.83	0.00	0.71	0.50	0.70	0.50	0.90	0.00
127	0.60	0.80	0.74	0.75	0.00	0.00	0.37	0.55	0.77	0.45	0.61	0.00
177	0.45	0.80	0.58	0.50	0.83	0.33	0.65	0.83	0.57	0.25	0.63	0.50
270	0.40	0.20	0.73	N/A	0.35	0.00	0.26	0.33	0.61	0.00	0.50	0.00
282	0.55	0.71	0.52	0.33	0.75	0.50	0.69	0.77	0.86	1.00	1.00	0.00

298	0.83	0.83	0.64	0.00	0.50	0.00	0.57	0.78	0.40	0.00	0.00	0.80
301	0.52	0.78	0.71	1.00	0.25	1.00	0.66	0.86	0.75	0.83	0.00	0.75
456	0.53	0.33	0.79	0.00	0.50	N/A	0.48	0.90	0.71	1.00	1.00	0.00
520	0.42	1.00	0.64	0.50	0.88	0.00	1.00	0.90	0.96	1.00	1.00	N/A
537	0.50	0.50	0.00	1.00	0.30	N/A	0.45	0.40	0.55	0.75	0.50	0.50
611	0.35	0.40	0.57	0.25	0.40	0.00	0.20	0.38	0.33	0.17	0.50	N/A
631	0.50	0.25	0.30	0.50	0.33	0.00	0.22	0.23	0.73	0.69	0.78	N/A
633	0.29	0.58	0.67	0.33	0.50	0.00	0.50	0.17	0.62	0.64	0.91	N/A
914	0.25	0.32	0.55	0.60	0.88	0.00	0.10	0.56	0.67	1.00	0.80	N/A
988	0.20	0.43	0.59	0.00	0.42	N/A	0.35	0.55	0.68	N/A	N/A	0.50

BIBLIOGRAPHY

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task? *Language Learning Journal, 39*(3), 269-280. doi:10.1080/09571736.2010.545.239

Alabbad, A., & Gitsaki, C. (2011). Attitudes toward learning English: A case study of university students in Saudi Arabia. In C. Gitsaki (Ed.), *Teaching and Learning in the Arab World* (pp. 3-28). Bern, IN: Peter Lang.

Andersen, R. W. (1976). *A functor acquisition study in Puerto Rico*. Paper presented at the 10th annual meeting of the Teachers of English to Speakers of Other Languages, New York.

Andersen, R. W. (1977). The impoverished state of cross-sectional morpheme acquisition accuracy methodology. In C. Henning (Ed.), *Proceedings of the Second Language Research Forum* (pp. 308-319). Los Angeles: University of California, Department of Applied Linguistics.

Andersen, R. W. (1978). An implicational model for second language research. *Language Learning, 28*(2), 221-282. doi:10.1111/j.1467-1770.1978.tb00134.x

Bailey, N., Madden, C., and Krashen, S. (1974). Is there a "natural sequence" in adult second language learning? *Language Learning, 24*(2), 235-343. doi:10.1111/j.1467-1770.1974.tb00505.x.

Ball, J. (1996). *Age and natural order in second language acquisition.* (Doctoral dissertation). University of Rochester, Rochester, NY.

Bardovi-Harlig, K. (1992). A second-look at t-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390-395. doi:10.2307/3587016

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Oxford: Blackwell.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33*(1), 1-17. doi:10.1111/j.1467-1770.1983.tb00983.x

Brown, J. D. (1983). An exploration of morpheme-group interactions. In K. Bailey, M. Long &

S. Peck (Eds.), *Second language acquisition studies* (pp. 25-40). Rowley, MA: Newbury House.

Brown, R. (1973). *A First Language: the Early Stages*. Cambridge, MA: Harvard University Press.

Brumfit, C. J. (1984). *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy*. Cambridge: Cambridge University Press.

Cadierno, T. (2008). Learning to talk about motion in a second language. In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 239-275). New York: Routledge.

Chang, J. (2001). Chinese Speakers. In M. Swan & B. Smith (Eds.), *Learner English: a teacher's guide to interference and other problems* (2nd ed., pp. 310-325). Cambridge: Cambridge University Press.

Chen, J. F., Warden, C. A., & Chang, H. T. (2005). Motivators that do not motivate: The case of Chinese EFL learners and the influence of culture on motivation. *TESOL Quarterly, 39*(4), 609-633. doi:10.2307/3588524

Cook, V. (1993). *Linguistics and second language acquisition*. New York: St. Martin's Press.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics 5*, 161-170. doi:10.1515/iral.1967.5.1-4.161

Corrigan, A., Dobson, B., Kellerman, E., Spaan, M., Strowe, L., & Tyma, S. (1979). *Michigan Test of English Language Proficiency (Form Q)*. Ann Arbor: Michigan University Press.

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics, 11*(2), 183-199. doi:10.1093/applin/11.2.183

de Bot, K. (2008). Introduction: Second language development as a dynamic process. *The Modern Language Journal, 92*(2), 166-178. doi:10.1111/j.1540-4781.2008.00712.x

De Villiers, J., & De Villiers, P. (1973). A cross-sectional study of the development of grammatical morphemes in child speech. *Journal of Psycholinguistic Research, 2*(3), 267-278. doi:10.1007/BF01067106

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9-42). Bristol, UK: Multilingual Matters.

Dulay, H., & Burt, M. (1973). Should we teach children syntax? *Language Learning, 23*(2), 245-258. doi:10.1111/j.1467-1770.1973.tb00659.x

Dulay, H., & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning, 24*(10), 37-53. doi:10.1111/j.1467-1770.1974.tb00234.x

Dulay, H., Burt, M., & Krashen, S. (1982). *Language Two*. New York: Oxford University Press.

Eisenstein, M., & Starbuck, R. (1989). The effect of emotional investment on L2 production. In S. Gass, C. Madden, D. Preston & L. Selinker (Eds.), *Variation in Seond Language Acquisition: Psycholinguistic Issues* (pp. 125-137). Clevedon, UK: Multilingual Matters.

Ellis, N. (2006). Selective attention and transfer phenomena in L2 acquisition: contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*(2), 164-194. doi:10.1093/applin/aml015

Ellis, N., & Larsen-Freeman, D. (2009). *Language as a Complex Adaptive System*. Ann Arbor, MI: Language Learning Research Club.

Ellis, R. (1987). Interlanguage variability in narrative discourse: style shifting in the past tense. *Studies in Second Language Acquisition, 9*(1), 12-20. doi:10.1017/S0272263100006483

Ellis, R. (2003). *Task-Based Language Learning and Teaching*. Oxford: Oxford University Press.

Ellis, R. (2008). *The Study of Second Language Acquisition* (3rd ed.). Oxford: Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics, 30*(4), 474-509. doi:10.1093/applin/amp042

Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford: Oxford University Press.

English Language Institute at the University of Pittsburgh. (2007). RSA Evaluation and Rubric. Pittsburgh: English Language Institute.

Fathman, A. (1975). Language background, age and the order of acquisition of English structures. In M. Burt & H. Dulay (Eds.), *On TESOL '75: New directions in second language learning, teaching and bilingual education* (pp. 33-43). Washington, DC: TESOL.

Fen-Chuan Lu, C. (2001). The acquisition of English articles by Chinese learners. *Second Language Studies, 20*(1), 43-78.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task based learning. *Studies in Second Language Acquisition, 18*(3), 299-324. doi:10.1017/S0272263100015047

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375. doi:10.1093/applin/21.3.354

Fuller, J. K. (1978). *An investigation of natural and monitored difficulty orders by non-native adult speakers of English.* (Doctoral dissertation). Florida State University.

Gass, S. M., & Selinker, L. (2008). *Second Language Acquisition: An Introductory Course* (3rd ed.). New York, NY: Routledge.

Goldschneider, J. M., & DeKeyser, R. M. . (2001). Explaining the "Natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 51*(1), 1-50. doi:10.1111/1467-9922.00147

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139-150.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning, 2*(2)*4*, 287-298. doi:10.1111/j.1467-1770.1974.tb00509.x

Hakuta, K. (1976). A case study of a Japanese child learning English as a second language. *Language Learning, 26*(2), 321-351. doi:10.1111/j.1467-1770.1976.tb00280.x

Hakuta, K., & Cancino, H. (1977). Trends in second language acquisition research. *Harvard Educational Review, 47*, 294-316. Retrieved from http://her.hepg.org/content/e03v062m64745872/

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power.* London: Palmer Press.

Hatch, E., & Farhady, H. (1982). *Research design and statistics for applied linguistics*. Rowley, MA: Newbury House.

Hawkins, R., Al-Eid, S., Almahboob, I., Athanasopoulos, P., Chaengchenkit, R., Hu, J., Rezai, M., Jaensch, C., Jeon, Y., Jiang, A., Leung, Y-K. I., Matsunaga, K., Ortega, M., Sarko, G., Snape, N., & Velasco-Zárate, K. (2006). Accounting for English article interpretation by L2 speakers. In S. H. Foster-Cohen, M. M. Krajnovic & J. M. Djigunović (Eds.), *EUROSLA Yearbook 6* (pp. 7-25). Amsterdam: John Benjamins.

Houck, N., Robertson, J., & Krashen, S. (1978). On the domain of the conscious grammar: Morpheme orders for corrected and uncorrected ESL student transcriptions. *TESOL Quarterly, 12*, p. 335-339.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473. doi:10.1093/applin/amp048

Howell, D. C. (2002). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Huang, J., Li, A., & Li, Y. (2009). *The Syntax of Chinese*. New York: Cambridge University Press.

Hunt, K. (1970). *Syntactic maturity in school-children and ddults*: Monograph of the Society for Research into Child Development.

Izumi, E., & Isahara, H. (2004). Investigation into language learners' acquisition order based on an error analysis of a learner corpus. *IWLeL: An interaction Workshop on Language e-Learning*, 63-71.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *IRAL: International Review of Applied Linguistics in Language Teaching, 28*(2), 99-118. doi:10.1515/iral.1990.28.2.99,

Juffs, A., & Friedline, B. (2014). Sociocultural influences on the use of a web-based tool for learing English vocabulary. *System, 42*, 48-59. doi:10.1016/j.system.2013.10.015

Kamimoto, T., Shimura, A., & Kellerman, E. (1992). A second language classic reconsidered-the case of Schachter's avoidance. *Second Language Research, 8*(3), 251-277. doi:10.1177/026765839200800305

Kellerman, E., & Sharwood-Smith, M. (1986). *Crosslinguistic influence in second language acquisition*. New York: Pergamon.

Kessler, C., & Idar, A. (1979). Acquisition of English by a Vietnamese child. *Working Papers on Bilingualism, 18,* 66-79.

Kjaarsgard, M. (1979). *The order of English morpheme category acquisition by Vietnamese children.* (Doctoral dissertation). Arizona State University, Phoenix, AZ.

Koike, I. (1983). *Acquisition of grammatical structures and relevant verbal strategies in a second language*. Tokyo: Taishyukan.

Krashen, S. (1977). Some issues relating to the Monitor Model. In C. Y. H. Brown, & R. Crymes (Eds.), *On TESOL '77*. Washington, DC: TESOL.

Krashen, S. (1985). *The Input Hypothesis: Issues and Implications*. New York: Longman.

Krashen, S., Bailey, N., & Madden, C. (1975). Theoretical aspects of grammatical sequencing. In M. Burt & H. Dulay (Eds.), *On TESOL '75: New directions in second language learning, teaching, and bilingual education* (pp. 44-54). Washington, DC: TESOL.

Krashen, S., Butler, J., Birnbaum, R., & Robertson, J. (1978). Two studies in language acquisition and language learning. *ITL, Review of the Institute of Applied Linguistics Louvain*, (39-40), 73-92.

Krashen, S., Houck, N., Giunchi, P., Bode, S., Birnbaum, R., & Strei, G. (1977). Difficulty order for grammatical morphemes for adult second language performers using free speech. *TESOL Quarterly*, *11*(3), 338-341.

Krashen, S., Sferlazza, V., Feldman, L., & Fathman, A. (1976). Adult performance on the SLOPE test: More evidence for a natural sequence in adult second language acquisition. *Language Learning, 26,* 145-151.

Lardiere, D. (2007). *Ultimate attainment in second language acquisition: a case study*. Mahwah, NJ: Lawrence Earlbaum Associates.

Lardiere, D. (2009). Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research, 25*(2), 173-227. doi:10/1177/0267658308100283

Larsen-Freeman, D. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly*, *9*(4), 409-430. Retrieved from http://www.jstor.org/stable/3585625

Larsen-Freeman, D. (1976). An explanation for the morpheme acquisition order of second language learners. *Language Learning, 26*(1), 125-134. doi:10.1111/j.1467-1770.1976.tb00264.x

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18*(2), 141-165. doi:10.1093/applin/18.2.141

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English *Applied Linguistics, 27*(4), 590-619. doi:10.1093/applin/aml029

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579-589. doi:10.1093/applin/amp043

Larsen-Freeman, D., & Long, M. (1991). *An introduction to second language acquisition research*. London: Longman.

Larson-Hall, Jenifer. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

Lennon, P. (1991). Error: some problems of definition, identification and distinction. *Applied Linguistics, 12*(2), 180-195. doi:10.1093/applin/12.2.180

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Ann Arbor: The University of Michigan Press.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Li, Y-H. A. (1999). Plurality in a classifier language. *Journal of East Asian Linguistics, 8*(1), 75-99. doi:10.1023/A:1008306431442

Lightbown, P. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. Seliger & M. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 217-243). Rowley, MA: Newbury House.

Lightbown, P., Spada, N., & Wallace, R. (1980). Some effects of instruction on child and adolescent ESL learners. In R. Scarcella & S. Krashen (Eds.), *Research in second language acquisition* (pp. 162-172). Rowley, MA: Newbury House.

Lin, J-W. (2006). Time in a language without tense: The case of Chinese. *Journal of Semantics, 23*(1), 1-53. doi:10.1093/jos/ffh033

Long, M., & Sato, C. (1984). Methodological issues in interlanguage studies: an interactionist approach. In C. C. A. Davies, A. R. P. Howatt (Eds.), *Interlanguage*. Edinburgh, UK: Edinburgh University Press.

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence From the acquisition of plural - s, articles, and possessive 's. *Language Learning, 59*(4), 721-754. doi:10.1111/j.1467-9922.2009.00524.x

Luoma, S. (2004). *Assessing Speaking*. New York: Cambridge University Press.

Mace-Matluck, B. (1977). *The order of acquisition of certain oral English structures by native-speaking children of Spanish, Cantonese, Tagalog, and Ilokano learning English as a second language between the ages of five and ten.* (Doctoral dissertation). University of Texas at Austin.

Mace-Matluck, B. (1979). The order of acquisition of English structures by Spanish speaking children: Some possible determinants. In R. W. Anderson (Ed.), *The acquisition and use of Spanish and English as first and second languages* (pp. 75-89). Washington, DC: TESOL.

Makino, T. (1979). English morpheme acquisition order of Japanese secondary school students. *TESOL Quarterly, 13*(3), 428-449.

Malvin, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. W. A. Ryan (Ed.), *Evolving Models of Language* (pp. 58-71). Clevedon: Multilingual Matters.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System,*

25(2), 215-232. doi:10.1016/S0346-251X(97)00010-9.

McCormick, D. E., & Vercellotti, M. L. (2009). To err is human to self-correct divine: Examining classroom recorded speaking activity data to support ESL self-correction as noticing. *Paper Presented at AAAL*. Denver, CO.

Morgan-Short, K., Finger, I., Grey, S., & Ullman, M. T. (2012). Second language processing shows increased native-like neural responses after months of no exposure. *PLoS ONE, 7*(3), 1-18. doi: 10.1371/journal.pone.0032974

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. doi:10.1093/applin/amp044

Nuibe, Y. (1986). A report on the development of the grammatical morphemes in Japanese junior high school students learning English as a foreign language. *Kyoiku Kagaku* [Educational Science], *28*(2), 371-381.

Nydell, M. K. (2012). *Understanding Arabs: A guide for modern times* (5th ed.). Boston: Intercultural Press, Inc.

O'Brien, L., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29*(4), 557-581. doi: 10.10170S/27226310707043X

Odlin, T. (1989). *Language transfer: Cross-Linguistic influence in language learning*. Cambridge: Cambridge University Press.

Pak, Y. (1987). *Age differences in morpheme acquisition among Korean ESL learners: Acquisition order and acquisition rate.* (Doctoral dissertation). University of Texas at Austin.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601. doi:10.1093/applin/amp045

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning, 33*(4), 65-497. doi:10.1111/j.1467-1770.1983.tb00945.x

Pienemann, M. (1998). *Language Processing and Second Language Development*. Amsterdam: John Benjamins.

Pienemann, M., & Johnston, M. (1985). *Towards an explanatory model of language acquisition*. Paper presented at the Second Language Research Forum, University of California at Los Angeles.

Po-Ching, Y., & Rimmington, D. (2004). *Chinese: A Comprehensive Grammar*. London:

Routledge.

Poeppel, D. and Wexler, K. (1993). The full competence hypothesis of clause structure in early German. *Language 69*(1), 1-32. Retrieved from http://www.jstor.org/stable/416414

Porter, J. (1977). A cross-sectional study of morpheme acquisition in first language learners. *Language Learning, 27*(1), 47-62. doi:10.1111/j.1467-1770.1977.tb00291.x

Qafisheh, H. A. (1977). *A Short Reference Grammar of Gulf Arabic*. Tucson, AZ: The University of Arizona Press.

Ribes, K. Z. (2012). World Arabic Language Day. *UNESCO.* Retrieved December 23, 2013, from http://unesdoc.unesco.org/images/0021/002191/219174E.pdf

Rickford, J. R. (2002). Implicational scales. In J. K. Chambers, P. Trudgill & N. Schillings-Estes (Eds.), *The Handbook of language variation and change* (pp. 142-167). Malden, MA: Blackwell Publishing.

Riddle, P. (1993). *The relationship of the bilingual syntax measure to the English morpheme development measure.* (Doctoral dissertation). Temple University, Philadelphia, PA.

Robinson, P. (2003). Attention and memory during SLA. In M. Long & C. J. Doughty (Eds.), *The Handbook of Second Language Acquisition* (pp. 631-678). Malden, MA: Blackwell Publishing.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43*(1), 1-32. doi:0019042X/2005/043-01

Robinson, P., Cardierno, T., & Shirai, Y. (2009). Time and Motion: Measuring the Effects of the Conceptual Demands of Tasks on Second Language Speech Production. *Applied Linguistics, 30*(4), 533-554. doi:10.1093/applin/amp046

Rosaldo, L. (1986). *The role of the Monitor in the acquisition sequence of twelve English morphemes by adult English as a second language learners from four different language groups.* (Doctoral dissertation). Texas A & I University, Kingsville, TX.

Rosansky, E. (1976). Methods and morphemes in second language acquisition research. *Language Learning, 26*, 409-425. doi:10.1111/j.1467-1770.1976.tb00284.x

Sasaki, M. (1987). Is Uguisu an exceptional case of “idiosyncratic variation”? Another counterexample to the “natural order”. *Chugoku-Shikoku Academic Society of Education Research Bulletin, 32*, 170-174.

Schepps, H. (2013). *The role of language background (Arabic vs. Chinese), context, and communicative orientation in English interlanguage plural morphology*. Unpublished manuscript. Department of Linguistics. University of Pittsburgh.

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research, 12*(1), 40-72. doi: 10.1177/026765839601200103

Segalowitz, N. (2007). Access fluidity, attention control, and the acquisition of fluency in a second language. *TESOL Quarterly, 41*(1), 181-186. doi: 10.1002/j.1545-7249.2007.tb00047.x

Shin, S., & Milroy, L. (1999). Bilingual language acquisition by Korean school children in New York City. *Bilingualism: Language and Cognition, 2*, 147-167.

Shirahata, T. (1988). The learning order of English grammatical morphemes by Japanese high school students. *JACET Bulletin, 19*, 83-102.

Shirai, Y. (1992). Conditions on transfer: A connectionist approach. *Issues in Applied Linguistics, 3*, 91-120.

Shirai, Y. (2002). The prototype hypothesis of tense-aspect acquisition in second language. In R. Salaberry & Y. Shirai (Eds.), *The L2 Acquisition of Tense-aspect Morphology* (pp. 455-478). Philadelphia: John Benjamins.

Skehan, P. (1989). *Individual differences in second-language learning*. London: Edward Arnold.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1-14. doi:10.1017/S026144480200188X

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532. doi:10.1093/applin/amp047

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-211. doi:10.1177/136216889700100302

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120. doi:10.1111/1467-9922.00071

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In S. V. Daele, A. Housen, F. Kuiken, M. Pierrard & I. Vedder (Eds.), *Complexity, Accuracy, and Fluency in Second Language Use, Learning, and Teaching*. Brussels: University of Brussels Press.

Smith, B. (2001). Arabic Speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 195-213). Cambridge: Cambridge

University Press.

Spinner, P. (2007). *Placement Testing and Morphosyntactic Development in Second Language Learners of English.* (Doctoral dissertation). University of Pittsburgh.

Spinner, P. (2011). Second language assessment and morphosyntactic development. *Studies in Second Language Acquisition, 33*(4), 529-561. doi:10.1017/S0272263111000301

Stockwell, R., Bowen, J., & Martin, J. (1965). *The grammatical structures of English and Spanish*. Chicago: University of Chicago Press.

Tarone, E. E. (1985). Variability in interlanguage use: A study of style shifting in morphology and syntax. *Language Learning, 35*, 373-403. doi:10.1111/j.1467-1770.1985.tb01083.x

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 239-276). Amsterdam: John Benjamins.

Thelen, E., & Bates, E. (2003). Connectionism and dynamic systems: Are they really different? *Developmental Science, 6*(4), 378-391. doi:10.1111/1467-7687.00294

Thewissen, J. (2013). Capturing L2 Accuracy Developmental Patterns: Insights From an Error-Tagged EFL Learner Corpus. *The Modern Language Journal, 97*(S1), 77-101. doi:10.1111/j.1540-4781.2012.01422.x

Ullman, M. T. (2001). The Declarative/Procedural model of lexicon and grammar. *Journal of Psycholinguistic Research, 30*(1), 37-69. doi:10.1023/A:1005204207369

VanPatten, B. (2007). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 115-135). Mahwah, NJ: Lawrence Erlbaum Associates.

Vercellotti, M. L. (2012). *Complexity, Accuracy, and Fluency as Properties of Language Performance: The Development of the Multiple Subsystems over Time and in Relation to Each Other.* (Ph. D.), University of Pittsburgh.

Verspoor, M., Lowie, W., & de Bot, K. (2007). Input and Second Language Development from a Dynamic Perspective. In T. Y.-S. Piske, M. (Ed.), *Input Matters in SLA* (pp. 62-80). Bristol, UK: Multilingual Matters.

White, Lydia. (1989). *Universal grammar and second language acquisition* (Vol. 1). Philadelphia: John Benjamins Publishing.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i: Second Language Teaching and Curriculum Center.

Young, R. (1993). Functional constraints on variation in interlanguage morphology. *Applied Linguistics, 14*(1), 76-97. doi:10.1093/applin/14.1.76

Young-Scholten, M., Ijuin, C., & Vainikka, A. (2005). *Organic grammar as a measurement of development*. Paper presented at the 39th Annual TESOL Conference, San Antonio, TX.

Zobl, H., & Liceras, J. (1994). Functional categories and acquisition orders. *Language Learning, 44*(1), 159-180. doi:10.1111/j.1467-1770.1994.tb01452.x